JCMM

Journal of Computers, Mechanical and Management

e-ISSN: 3009-075X

Volume 2, Issue 6

2023

AAN PUBLISHING

Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AAN PUBLISHING

Editorial Comments: JCMM Volume 2 Issue 6

Nanjangud Subbaro Mohan*[1] and Lokesh Choudhary†[2]

[1]Journal of Computers, Mechanical and Management, AAN Publishing, Kangar Perlis, Malaysia 01000
[2]Department of Multidisciplinary Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

This issue of the *Journal of Computers, Mechanical and Management* (JCMM) showcases six articles spanning traffic noise modeling, sentiment analysis, breast cancer detection, fundus image generation, masonry repair materials, and RFID technology in shopping carts. This issue showcases selected and extended versions of exceptional research papers originally presented at the International Conference on Smart and AI-enabled Technology for Sustainable Development (SAIT for SD 2023), held on September 12-13, 2023. The conference brought together scholars, researchers, and experts from diverse fields to discuss and explore the latest advancements in technology and their applications for sustainable development. The research papers included in this volume reflect a rich diversity of topics, ranging from cutting-edge developments in computer science to innovative solutions in core engineering fields and strategic management. Each paper demonstrates a commitment to advancing knowledge and promoting sustainable practices..

The first article, Traffic Noise Prediction for Delhi-NCR Using Multiple Regression Modelling Approach by Mann and Singh, investigates traffic noise pollution in Delhi. The authors developed a regression model using data from 31 locations to identify significant variables such as road geometry, traffic volume, and environmental factors affecting noise levels. The model achieved a coefficient of determination, R^2, of 0.75, indicating a robust predictive capability [1].

Saggu, Pal, and Dev contribute with Bond Strength of Substrate With Repair Material for Masonry Structures, which presents a scientometric analysis of masonry repair materials. Using CiteSpace, they identified research trends and gaps in the durability and compatibility of repair materials with historic masonry [2].

In Enhanced Shopping Experiences: The Role of RFID Technology in Smart Carts, Datta et al. explore the application of RFID technology to improve retail experiences. They designed a smart cart system that automatically calculates costs, streamlining the checkout process in supermarkets [3].

Jain and Singla's Breast Cancer Detection using Machine Learning Algorithms evaluates various ML algorithms, including SVM and Decision Trees, for classifying breast cancer data. The study highlights the efficiency of ML methods in enhancing diagnostic accuracy, with SVM emerging as the most effective [4].

Sentiment Analysis on IMDB Review Dataset by Singh and Singla employs deep learning models such as BiLSTM to analyze movie reviews. This research underscores the potential of BiLSTM in achieving high accuracy in sentiment classification tasks [5]. Finally, Kapoor and Arora present Fundus Image Generation using EyeGAN, which proposes an enhanced GAN model for medical image augmentation. Their EyeGAN model generated synthetic fundus images, improving dataset balance for retinal disease detection [6].

This issue exemplifies JCMM's commitment to publishing high-quality research that contributes to technological advancements and interdisciplinary applications.

*Editor-in-Chief: editor@jcmm.co.in
†Guest Editor:lokeshchoudhary@ncuindia.edu

Published: 03 January 2024

DOI: 10.57159/gadl.jcmm.2.6.230112.

References

[1] Mann, S., & Singh, G. (2023). Traffic Noise Prediction for Delhi-NCR Using Multiple Regression Modelling Approach. *Journal of Computers, Mechanical and Management*, *2*(6). doi:10.57159/gadl.jcmm.2.6.230104.

[2] Saggu, K., Pal, S., & Dev, N. (2023). Bond Strength of Substrate With Repair Material for Masonry Structures. *Journal of Computers, Mechanical and Management*, *2*(6). doi:10.57159/gadl.jcmm.2.6.230107.

[3] Datta, I., Garg, A., Tonk, A., & Rakheja, P. (2023). Enhanced Shopping Experiences: The Role of RFID Technology in Smart Carts. *Journal of Computers, Mechanical and Management*, *2*(6). doi:10.57159/gadl.jcmm.2.6.230110.

[4] Jain, B., & Singla, N. (2023). Breast Cancer Detection using Machine Learning Algorithms. *Journal of Computers, Mechanical and Management*, *2*(6). doi:10.57159/gadl.jcmm.2.6.230109.

[5] Singh, S. K., & Singla, N. (2023). Sentiment Analysis on IMDB Review Dataset. *Journal of Computers, Mechanical and Management*, *2*(6). doi:10.57159/gadl.jcmm.2.6.230108.

[6] Kapoor, P., & Arora, S. (2023). Fundus Image Generation using EyeGAN. *Journal of Computers, Mechanical and Management*, *2*(6). doi:10.57159/gadl.jcmm.2.6.230106.

Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 2 Issue 6

Article Number: 230104

Traffic Noise Prediction for Delhi-NCR Using Multiple Regression Modelling Approach

Suman Mann*and Gyanendra Singh

Department of Civil Engineering, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Haryana, India 131039

Abstract

Traffic noise prediction models are crucial for designing highways to implement preventive measures against traffic noise pollution by analyzing future trends. This study aims to identify the traffic, road geometrical, and environmental parameters that escalate traffic noise pollution, enabling rectification of influencing factors and enhancement of strategies to reduce this pollution. A traffic noise prediction model was developed for the highways of Delhi-NCR using the Multiple Regression approach, incorporating various traffic, geometric, and environmental parameters. Statistical analysis was conducted, and the model was formulated based on data collected from 31 sampling stations on two major Delhi highways. Significant variables identified include the number of lanes, average building height, international roughness index, temperature, wind speed, and humidity. The model's validity is affirmed by a coefficient of determination R2 = 0.75, indicating a good fit.

Keywords: Traffic Noise Prediction Model; Noise Pollution; Significant Variables; Multiple Regression Modelling

1 Introduction

Traffic noise is an important factor to be considered in relation with public health [1]. It is rising as a threat to environment, causing health issues to people living in neighborhood of roads and reducing their quality of life [2]. High levels of noise lead to stress reactions in human body, that occurs even during sleep, and these stress reactions lead to hypertension, cardiovascular disease, cognitive impairment, annoyance and can lead to premature deaths. Traffic noise problem is associated with urban and roadway infrastructure developments and technological progress. Traffic noise management can be achieved by proper planning, designing, traffic control and police measures. Urban planners rely on Traffic Noise Prediction (TNP) models in strategizing noise mitigation measures [3]. Noise models involves assessing and predicting traffic noise levels based on various field measurements [4]. Traffic volume, traffic composition, traffic speed and road geometry are the most significant variables in predicting traffic noise [4–6]. FHWA- USA, CORTN-UK, CNOSSOS- EU, ASJ-RTN, RLS-90 are conventional models used by different countries for traffic noise modelling [3, 5, 7–9]. Adjustment for intersections modelling were included in FHWA, RLS- 90, ASJ-RTN and CNOSSOS, whereas CORTN was later modified for noise estimation at intersections [8, 10, 11]. These models are based on homogenous traffic conditions, whereas countries like India face heterogenous traffic conditions, hence models developed for homogenous traffic conditions cannot be opted for heterogenous traffic conditions [12]. Indian research work in the field of traffic noise is very narrow in comparison to other developed countries [13]. Monazzam et al. developed TNP model using 9 different traffic and road related variables for Ahvaz city [14].

*Corresponding author: suman.mann@dseu.ac.in
Received: 14 November 2023; **Revised:** 30 November 2023; **Accepted**: 04 December 2023; **Published:** 31 December 2023

DOI: 10.57159/gadl.jcmm.2.6.230104.

Ranpise et al. developed an artificial neural network model for 3 major arterial roads of Surat, India [15]. Gharibi et al. gave a regression model for a highway in Golestan national park, Iran [16]. Suthanaya carried out traffic noise monitoring for Depanser city, Indonesia and modelled traffic noise against traffic composition [17]. Various TNP models have been developed using various modelling approaches (linear regression and machine learning), which include traffic related variables [9, 18–23]. Machine learning models require large data sets for non-linear mapping and get the benefit of generalization. Machine learning models predict better but cannot quantify the effect of variables contributing to noise pollution. Whereas regression models require small data sets and are able to quantify the effect of variables on traffic noise. Machine learning models are suitable for cost estimation due to traffic noise, but during the planning stage, regression models can be used to study the effect of variables so that proper mitigation measures can be provided. Various regression models have been developed but very few have comprehensively explored the effect of environmental factors along with traffic factors on traffic noise levels. Therefore, in the present study, a statistical model for predicting *Leq* is developed for heterogenous traffic conditions considering 12 different variables from traffic related factors, road related factors, and environmental factors. So that the influencing factors may be rectified, and enhancement can be done for reducing the traffic noise pollution.

2 Methods

2.1 Study area and location

Delhi, the capital of India, is selected as the study area. Two major roads of Delhi, NH-9 and Ring Road, were selected for the study. In total, 31 sampling sites were selected on these two roads based on different land use patterns. Details of the study locations are given in Table 1.

Table 1: Details of Sampling Stations

Sampling Station ID	**Location**
SS1	Shalimar Bagh
SS2	Shakurpur Telephone Exchange
SS3	Punjabi Bagh
SS4	Raja Garden
SS5	Naraina Industrial Area
SS6	Dhaula Kuan Passage
SS7	New Moti Bagh
SS8	Hyatt Hotel
SS9	AIIMS
SS10	South Extension 2
SS11	Lajpat Nagar
SS12	Nizammudin
SS13	IP
SS14	ISBT
SS15	ITO
SS16	Azadpur
SS17	Tikri Border
SS18	Mundka Industrial Area
SS19	Nangloi
SS20	Peeragarhi
SS21	Bhopura Border
SS22	Saraswati Vihar
SS23	Haiderpur
SS24	Mukundpur
SS25	Wazirabad Mode
SS26	Signature Bridge
SS27	Khazoori
SS28	Bhajan Pura
SS29	Yamuna Vihar
SS30	Ashok Nagar
SS31	Mandoli

2.2 Noise measurement

Lutron sound level meter SL 4033SD was used for measurement. For measuring noise levels, the instrument was mounted at a height of 1.5m from the ground on a tripod stand and a distance of 1m from the road façade. Sound levels were monitored two times a day for 1 hour each time. The sampling plot of the study location is shown in Figure 1.

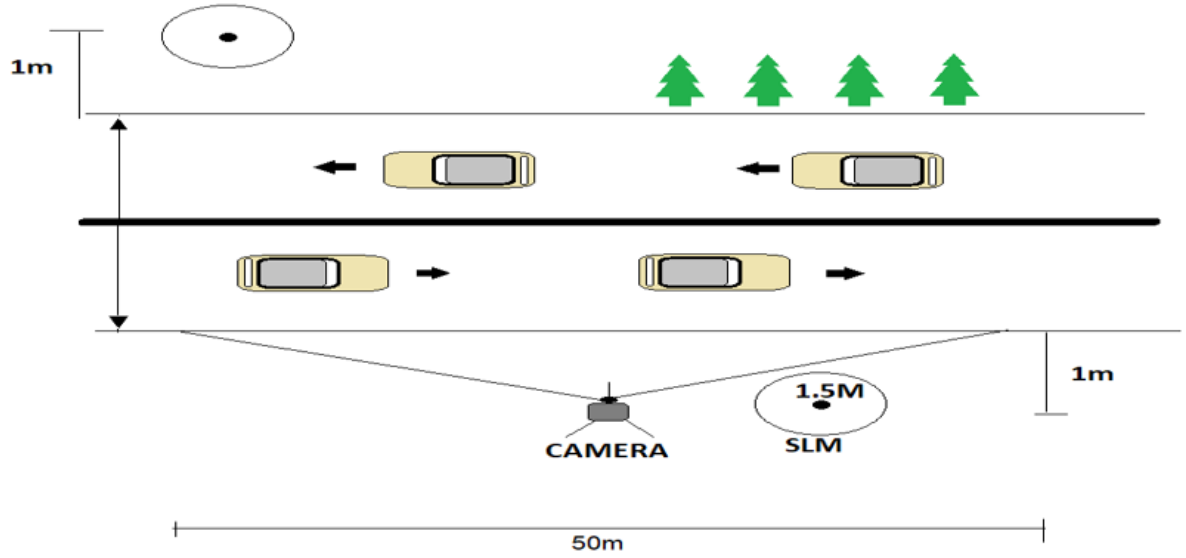


Figure 1: Graphical representation of site

2.3 Traffic, road and environmental data collection

Traffic data was obtained using a video graphic survey. Videos of traffic were simultaneously recorded along with noise measurements, and the videos were later used to obtain total traffic and traffic composition. At each specified location, a road marking of 15m was marked, and the time taken by vehicles to cover this distance was obtained by counting the number of frames. From this, the average traffic speed of the vehicles was calculated. The mobile app Road Bump Free was used to record the International Roughness Index (IRI), and other road-related data was obtained manually at the site. Meteorological data such as temperature, humidity, and wind speed were obtained from the AccuWeather mobile app. The average building height is determined by counting the number of floors and multiplying it by 3.5 m.

2.4 Model development

The whole data set of 31 stations was used to develop a model for Indian cities using regression analysis. Modelling based on *Leq*(1h) as the dependent variable and 12 other variables as independent variables (shown in Table 2) was performed using SPSS software. The effect of each independent variable on *Leq* was determined by using a scatter plot and linear regression.

Table 2: Model inputs (Independent variables affecting the sound)

Traffic factors	**Road factors**	**Location conditions**
Traffic flow (Veh/hr)	Roughness coefficient	Temperature
Average traffic speed and speed variance	Building height	Wind speed
Traffic composition in terms of percentage of 2W, 3W, 4W and/or HV		Humidity

3 Results and Discussion

3.1 Multiple linear regression modelling

The descriptive statistics of dependent and independent variables measured at two main roads of Delhi, on 31 sampling stations for 62 hours (one in rush hours and one in off rush hours) is presented in Table 3. Figure 2 represents the noise levels at all the sampling stations during rush hours and off rush hours. Measured noise levels are above the WHO recommended levels at all the sampling points, likely due to the business and commercial activities taking place in most locations.

Table 3: Descriptive statistics of variables

Variable	Observations	Minimum	Maximum	Mean	Std. Deviation
Dependent variable					
Leq 1 hr (dBA)	62	67.674	83.231	74.591	4.500
Independent variables					
Total traffic flow (vehicles/hr)	62	1807	7591	4721	1608.747
%2W	62	25.774	61.855	39.142	9.320
%3W	62	6.588	29.797	17.556	5.544
%4W	62	20.238	57.250	37.820	11.405
%HV	62	2.898	12.889	5.641	2.070
Average speed (KMPH)	62	24.555	52.874	40.527	5.872
Std deviation of speed (KMPH)	62	0.860	13.725	3.002	2.056
No of lanes (Numbers)	62	3	5	3.774	0.493
IRI (m/Km)	62	0.130	15.030	7.106	2.666
Avg building height (m)	62	0	17.5	7.903	6.010
Temperature (˚C)	62	20	30	25.323	2.373
Wind speed (km/h)	62	4	12	7.226	2.028
Humidity (%)	62	35	63	45.806	6.458

For the development of the traffic noise model, Pearson's correlation coefficient test was carried out for a 95% confidence interval to determine the variables having a substantial effect on the dependent variable *Leq* (dBA). Variables with $r \geq 0.2$ were used to linearly regress with *LAeq*. Total traffic flow and Average speed are converted into logarithmic form to attain a linear relationship. Traffic volume and speed are considered in logarithmic form by various researchers [21, 24]. The model was developed using Log traffic flow, Log average traffic speed, Number of lanes, Average building height, International Roughness Index (IRI), Wind speed, Temperature, and Humidity. Variables with significance < 0.05 were considered significant, and other variables were removed from the analysis. Therefore, the Number of lanes, IRI, Average building height, Temperature, Wind speed, and Humidity were considered, and all other variables were rejected from the analysis. As per Table 4, $R^2 = 0.751$, R^2 explains the total variation in the dependent variable *LAeq* due to independent variables, or the contribution of independent variables in the change of dependent variables. In the present model, $R^2 = 0.751$ indicates that our independent variable has a 75% change in the dependent variable. Table 5 represents the comparison of measured and predicted *LAeq* for the data.

Table 4: Regression results

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.866a	.751	.723	2.3668

Table 5: Comparison of measured and predicted *LAeq* for the data

Pair 1	Mean	Std. Deviation (dBA)	Std. Error Mean (dBA)
LAeq 1hr	74.596	4.500	.571
Predicted Value	74.591	3.898	.495

Analysis of variance (ANOVA) test was performed to check the variance among independent variables and *LAeq* (dependent variable). ANOVA results show that the P value is 0.000, which is less than 0.05, hence we can say that there is a significant relationship between *LAeq* (dependent variable) and independent variables. Table 6 shows the coefficients results. The Beta value for the variable temperature is negative, whereas for the Number of lanes, IRI, Average building height, wind speed, and humidity, Beta values are positive. Negative values indicate a negative relationship between the dependent and independent variable, meaning when the temperature (having negative Beta values) increases by 1˚C, dependent variable *LAeq* decreases by 0.252 dBA. These results are in confirmation with the findings of Sánchez-Fernández et al. [25]. At increasing temperature, the speed of sound increases, leading it to refract towards higher altitude and leading to lower noise levels at the receiver end. Positive Beta values indicate a positive relation between *LAeq* (dependent variable) and independent variables (Number of lanes, IRI, Average building height, wind speed, and humidity). As per results, noise levels increase by 0.113 dBA with an increase in the number of lanes. Similar results were given by Lu et al. [26], explaining that more lanes indicate more traffic demand leading to an increase in traffic volume, which increases the noise levels. When IRI increases by 1 m/km, *LAeq* increases by 0.090 dBA, these results are in accordance with the results given by Soedirdjo et al. [27]. Average building height also showed a positive relation with noise levels. Tandel and Macwan also published similar results [28]. Murugun et al. proved a linear relationship between noise levels and wind speed [29].

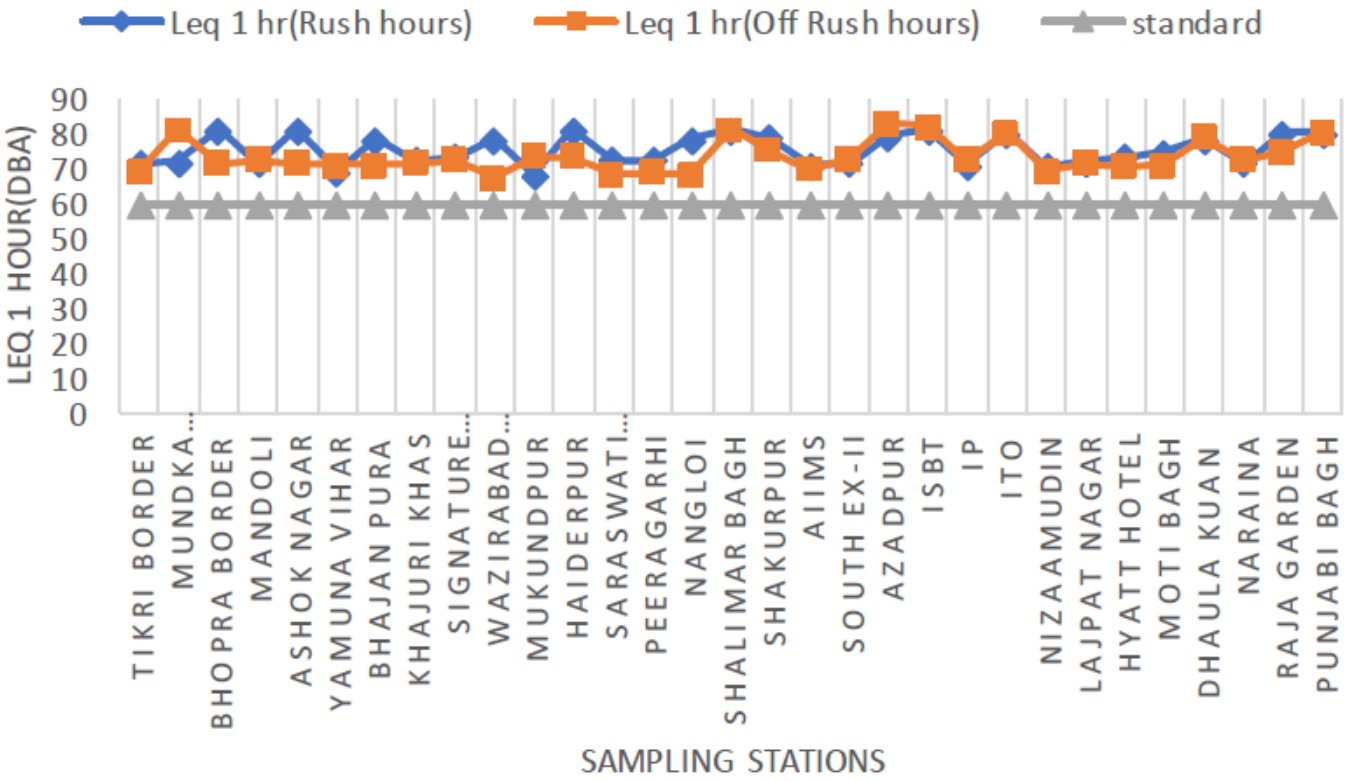


Figure 2: Noise levels at all sampling stations during rush hours and off rush hours

Table 6: Regression coefficients of the developed model

Model	Unstandardized Coefficients B	Standardized Coefficients Std. Error	Beta	t	Sig.
Constant	77.151	5.804		13.293	.000
No of lanes	1.033	.651	.113	1.587	.018
IRI	.152	.118	.090	1.287	.003
Avg building height	.107	.054	.143	2.000	.050
Temp	-.478	.191	-.252	-2.503	.015
Wind speed (km/h)	1.453	.209	.655	6.957	.000
Humidity (%)	.106	.051	.153	2.072	.043

4 Model validation

The accuracy of developed models was tested based on R^2 and other errors, as given in Table 7. The goodness of fit of the regression model was evaluated by comparing the observed and predicted values. A scatter plot between observed and predicted noise levels was plotted along a 45-degree line as shown in Figure 3. As per the developed model, the mean difference between predicted $LAeq$ 1hr and observed $LAeq$ 1hr is less than 0.0003 dBA. Thus, the developed model can predict traffic noise levels accurately for Indian cities.

Table 7: Comparison of measured and predicted $LAeq$ for the data

Pair 1	Mean	Std. Deviation (dBA)	Std. Error Mean (dBA)
$LAeq$ 1 hr	74.596	4.500	.571
Predicted Value	74.591	3.898	.495

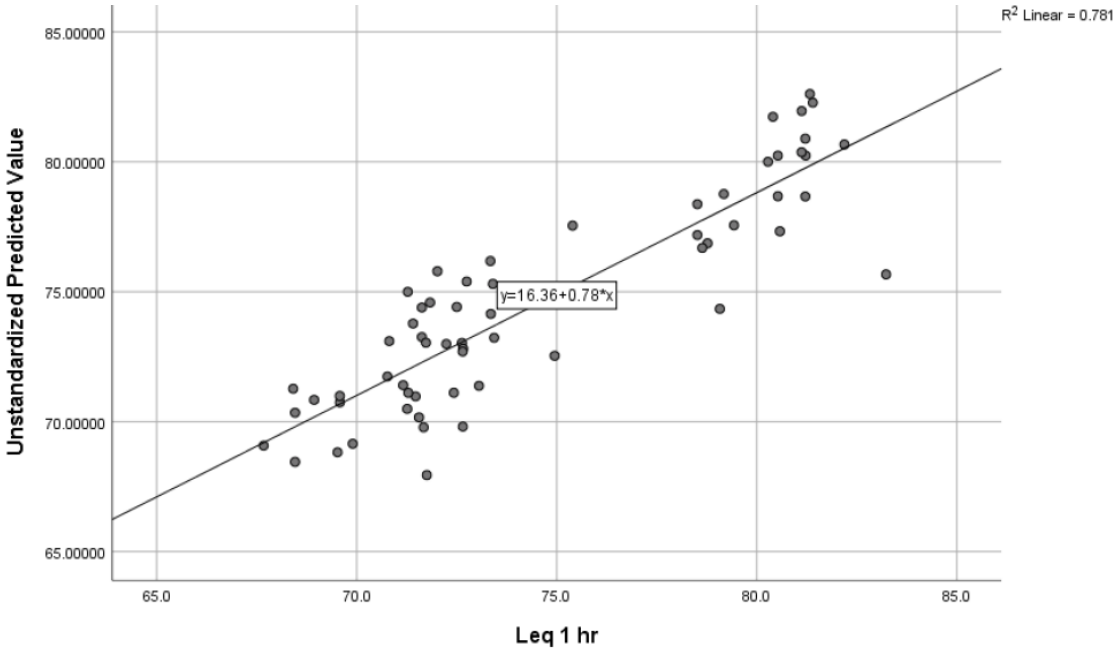


Figure 3: Scatter plot between predicted and observed *LAeq*

5 Conclusion

The present paper developed a traffic noise prediction model for highways in India with heterogeneous traffic conditions. Data sets for the dependent variable *LAeq* 1hr and twelve influencing variables were collected from two major roads in Delhi during May – June 2023. Out of the 12 explanatory variables, only 6 were found significant in the present study. The Number of lanes, IRI, Temperature, Humidity, Wind speed, and Average building height are considered in model development. The study revealed that IRI and Wind speed are the most significant factors affecting traffic noise. Results indicated that Temperature is negatively associated with noise levels, whereas all other factors are positively associated with noise levels. The mean difference between observed *LAeq* 1hr and predicted *LAeq* 1hr is 0.003 dBA. The developed model has higher prediction accuracy with a determination coefficient equal to 0.75. The developed regression model provides a simple linear equation which is easy to understand and can be used to check out the mitigation measures for reducing and controlling noise pollution due to traffic. The benefit of regression models is that they need fewer data sets and can be used in the planning stage where environmental impact assessment has to be done. The suggested mitigation measures in the study area are:

- Use of intelligent transport system to control traffic mobility, volume, speed, and composition.
- Technological solutions can be implemented: like the construction of noise barriers using sonic crystals, Poroelastic road surfaces for reduction of tyre/road noise.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

Suman Mann: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Investigation, Validation, writing; **Gyanendra Singh**: Supervision, Reviewing, validation and editing.

References

[1] M. Monazzam, V. Abolhasannejad, B. Moasheri, V. Abolhasannejad, and H. Kardanmoghaddam, "Noise pollution in old and new urban fabric with focus on traffic flow," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 35, no. 4, pp. 257–263, 2016.

[2] A. Yadav, M. Parida, and B. Kumar, "Statistical modeling of traffic noise at intersections in a mid-sized city, india," *Noise Mapping*, vol. 10, no. 1, p. 20220164, 2023.

[3] S. Sakamoto, "Road traffic noise prediction model 'asj rtn-model 2018': Report of the research committee on road traffic noise," *Acoustical Science and Technology*, vol. 41, no. 3, pp. 529–589, 2020.

[4] R. Makarewicz, M. Fujimoto, and P. Kokowski, "A model of interrupted road traffic noise," *Applied Acoustics*, vol. 57, no. 2, pp. 129–137, 1999.

[5] M. Delany, D. Harland, R. Hood, and W. Scholes, "The prediction of noise levels l10 due to road traffic," *Journal of Sound and Vibration*, vol. 48, no. 3, pp. 305–325, 1976.

[6] S. Abo-Qudais and A. Alhiary, "Statistical models for traffic noise at signalized intersections," *Building and Environment*, vol. 42, no. 8, pp. 2939–2948, 2007.

[7] T. Barry and J. Reagan, "Fhwa highway traffic noise prediction model," 1978.

[8] S. Kephalopoulos, M. Paviotti, and F. Anfosso-Lédée, *Common noise assessment methods in Europe. CNOSSOS-EU*. Luxembourg, 2012.

[9] C. Steele, "A critical review of some traffic noise prediction models," *Applied acoustics*, vol. 62, no. 3, pp. 271–287, 2001.

[10] N. Garg and S. Maji, "A critical review of principal traffic noise models: Strategies and implications," *Environmental Impact Assessment Review*, vol. 46, pp. 68–81, 2014.

[11] K. Abdur-Rouf and K. Shaaban, "Development of prediction models of transportation noise for roundabouts and signalized intersections," *Transportation Research Part D: Transport and Environment*, p. 103174, 2022.

[12] R. Kalaiselvi and A. Ramachandraiah, "A model for traffic noise prediction in heterogeneous traffic conditions," *Int J Curr Res*, vol. 4, pp. 180–184, 2012.

[13] K. Shalini and B. Kumar, "Development of traffic noise model (tnm) using regression analysis in varanasi city, india," *International Journal of Civil Engineering and Technology*, vol. 9, no. 4, pp. 70–76, 2018.

[14] E. Monazzam, M. Sekhavatjou, and A. Chabi, "Designing a traffic noise prediction model for highways in iranian megacities (case study: Ahvaz city)," vol. 8, no. 2, pp. 427–434, 2014.

[15] R. Ranpise, B. Tandel, and V. Singh, "Development of traffic noise prediction model for major arterial roads of tier-ii city of india (surat) using artificial neural network," *Noise Mapping*, vol. 8, no. 1, pp. 172–184, 2021.

[16] S. Gharibi, M. Aliakbari, A. Salmanmahiny, and H. Varastehe, "Evaluation and modelling of traffic noise on the asian highway in golestan national park, iran," in *MATEC Web of Conferences*, vol. 81, p. 04008, EDP Sciences, 2016.

[17] P. Suthanaya, "Modelling road traffic noise for collector road (case study of denpasar city)," *Procedia Engineering*, vol. 125, pp. 467–473, 2015.

[18] W. H. K. Lam and M. L. Tam, "Reliability analysis of traffic noise estimates in hong kong," *Transportation research Part-D*, vol. 3, no. 4, pp. 239–248, 1998.

[19] R. Stefano, D. Danato, and B. Morri, "A statistical model for predicting road traffic noise on poisson type traffic flow," *Noise control engineering journal*, vol. 49, no. 3, pp. 137–143, 2001.

[20] B. Li, T. Shu, and R. W. Dawson, "Evaluation and analysis of traffic noise from the main urban road in beijing," *Applied acoustics*, vol. 63, pp. 1137–1142, 2002.

[21] R. Golmohammadi, M. Abbaspour, P. Nassiri, and H. Mahjub, "Road traffic noise model," *Journal of Research in Health Sciences*, vol. 7, no. 1, pp. 13–17, 2007.

[22] O. Gundogdu, M. Gokdag, and F. Yuksel, "A traffic noise predication method based on vehicle composition using genetic algorithms," *Applied acoustics*, vol. 66, pp. 799–809, 2005.

[23] M. Tansatcha, P. Pamanikabud, A. Brown, and J. K. Affum, "Motorway noise modelling based on perpendicular propagation analysis of traffic noise," *Applied acoustics*, vol. 66, pp. 1135–1150, 2005.

[24] Z. P. Calixto A, Diniz FB, "The statistical modeling of road traffic noise in an urban setting," *Cities*, vol. 20, no. 1, pp. 23–29, 2003.

[25] M. Sánchez-Fernández, J. Morillas, D. González, and G. Gozalo, "Relationship between temperature and road traffic noise under actual conditions of continuous vehicle flow," *Transportation Research Part D: Transport and Environment*, p. 103056, 2021.

[26] X. Lu, J. Kang, P. Zhu, J. Cai, F. Guo, and Y. Zhang, "Influence of urban road characteristics on traffic noise," *Transportation Research Part D: Transport and Environment*, vol. 75, pp. 136–155, 2019.

[27] T. Soedirdjo, S. Hendarto, and A. Triadi, "Evaluation of traffic noise in arterial road with different road roughness," in *Proc. Eastern Asia Soc. Transportat. Studies*, vol. 4, pp. 1376–1389, 2003.

[28] B. Tandel, J. Macwan, and P. Ruparel, "Urban corridor noise pollution: a case study of surat city, india," in *Proceedings of International Conference on Environment and Industrial Innovation (ICEII 2011)*, vol. 12, pp. 144–148, June 2011.

[29] S. Murugan, V. Natarajan, and R. Kumar, "Noise model analysis and estimation of effect due to wind driven ambient noise in shallow water," *International Journal of Oceanography*.

Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 2 Issue 6

Article Number: 230106

Fundus Image Generation using EyeGAN: An Improved Generative Adversarial Network Model

Preeti Kapoor*[1] and Shaveta Arora[1]

[1]Department of Computer Science and Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

Abstract

Deep learning models are widely used in various computer vision fields ranging from classification, segmentation to identification, but these models suffer from the problem of overfitting. Diversifying and balancing the datasets is a solution to the primary problem. Generative Adversarial Networks (GANs) are unsupervised learning image generators which do not require any additional information. GANs generate realistic images and preserve the minute details from the original data. In this paper, a GAN model is proposed for fundus image generation to overcome the problem of labelled data insufficiency faced by researchers in detection and classification of various fundus diseases. The proposed model enriches and balances the studied datasets for improving the eye disease detection systems. EyeGAN is a nine-layered structure based on conditional GAN which generates unbiased, good quality, credible images and outperforms the existing GAN models by achieving the least Fréchet Inception Distance of 226.3. The public fundus datasets MESSIDOR I and MESSIDOR II are expanded by 1600 and 808 synthetic images respectively.

Keywords: Deep Learning; FID; Conditional GAN; Style GAN

1 Introduction

Image generation is the process of creating synthetic images which are realistic in nature [1]. Image augmentation is used in various computer vision problems to resolve the scarcity of original data and to overcome the problem of data imbalance. Various public fundus datasets also suffer from class imbalance and non-labelled data, and there is a lack of good quality images in large datasets. Though deep learning has recently gained popularity and outperformed many current methods in many fields, particularly when it comes to analyzing and reviewing fundus images, these models have poor generalizability and suffer from overfitting. Using an enriched and balanced dataset with labelled data for training can help in overcoming this problem. Well-trained deep learning models prove to provide better results and achieve good accuracy. There are various augmentation methods available for producing images like rotation, flipping, or cropping. With the increase in the use of neural networks in every field, these models are also used for image generation, producing real-like images. GAN models are widely used as augmentation techniques in several areas and also improve the overall results [1]. Lee et al. [2] suggested an any-time-of-the-day camera-based blind spot detection model: GAN + Active BSD system to fight the problem of nighttime sample availability. The cycle GAN was trained using publicly available Nexar data and was used to generate synthetic nighttime images by using the daytime side rectilinear images. Also, these images were used to annotate the synthetic nighttime images. While training the generator, the main aim was to find the mapping between day to night images. The data augmentation using generated samples doubled the samples at all 4 stages of the BSD system. Then all four stages were evaluated with and without data augmentation.

*Corresponding author: preeti19csd006@ncuindia.edu
Received: DD Month 20YY; **Revised:** DD Month 20YY; **Accepted**: DD Month 20YY; **Published:** DD Month 20YY

DOI: 10.57159/gadl.jcmm.2.6.230106.

The accuracy improved after nighttime data augmentation from 26% to 55%, which is two times in comparison to just day images. Liu et al. [3] proposed a GAN model for grape leaf disease images generation to overcome the problems faced in grape disease detection and called it Leaf GAN. The generator was recreated with digressive channel layers and discriminator with the compact connectivity approach and specified example normalization. For further stabilization of the training process, a deep regret analytic gradient penalty algorithm was used to train both generator and discriminator. Then identification was performed with and without using augmented data, also the proposed method LG, LG-RELU, LG-PRELU, was compared with existing DCGAN, WGAN. The proposed method achieved better performance as compared to other GANs and dataset without augmentation, with an efficiency improvement from 95.25% to 98.25% on Xception net. Also, it overcame the overfitting problem. Waheed et al. [4] proposed an auxiliary classifier GAN (ACGAN) and named it CovidGAN for generating artificial images. They then combined it with a CNN model to improve detection. For dimension reduction, Principal Component Analysis was used and confusion matrices were used for result visualization. An improved detection of COVID-19 with an accuracy of 95% was achieved. However, there was no control over the biased image generation since the class labels were not used. This paper focuses on producing good quality images for each class by creating a GAN model which incorporates labels as a condition and avoids biased image generation. The paper demonstrates the use of a GAN model as the augmentation technique [5] to enrich the selected DR Datasets and fight class inequality. A GAN model is presented which is based on conditional GAN for fundus image generation rather than just basic augmentation operations. The proposed method is made lighter than the base model by reducing the convolutional layers, i.e., a nine-layered architecture. Also, EyeGAN was compared with the existing GAN models using the FID score to highlight the better quality of images produced. MESSIDOR I and MESSIDOR II datasets are augmented using EyeGAN to balance the images in all the classes, thus making them enriched and balanced datasets. The major contributions of this paper are as follows:

- A GAN model is proposed which balances the target datasets and produces real-like, high-quality images, surpassing the results of existing models.
- The EyeGAN is compared with conditional GAN and Style GAN using the Fréchet Inception Distance (FID) score.
- Two enriched Diabetic Retinopathy (DR) Datasets are created, enhancing the capabilities for better detection and classification of eye diseases.

The whole work is presented in four sections. Section I is the survey of the various augmentation methods which are used for image generation. Section II explains the methodology, i.e., the structure of the proposed GAN model, the generator loss function, and datasets. Section III shows the comparison and the outcomes. Section IV concludes the research.

2 Data Augmentation

The image datasets are often captured under a certain set of circumstances such as in a different direction, position, scale, or brightness. Data augmentation exploits these characteristics to generate images. Figure 1 shows the various augmentation techniques used in computer vision.

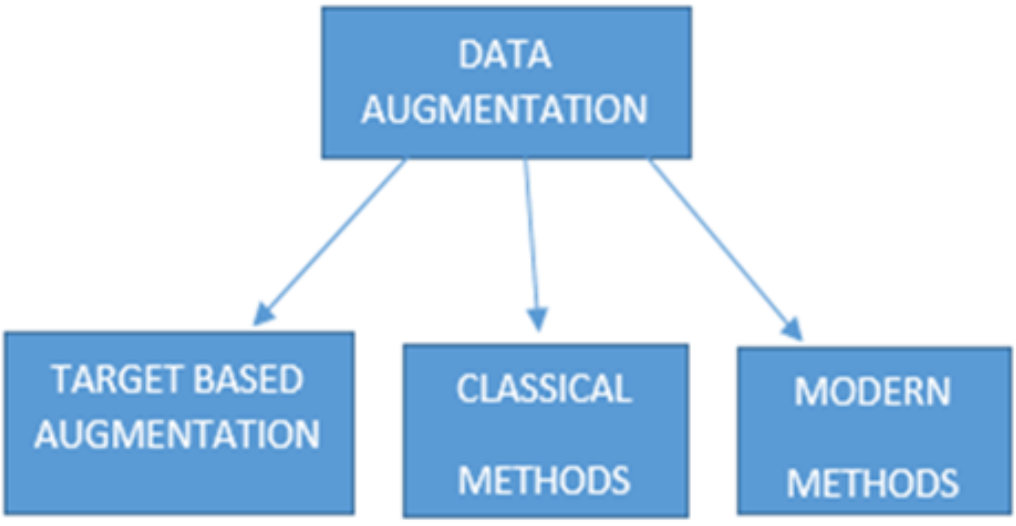


Figure 1: Various Augmentation techniques

2.1 Data augmentation based on target dataset size:

- **Offline augmentation:** This technique helps for small datasets, as the number of images are increased by a large factor. Here the whole dataset is targeted before it is fed to the machine learning models.
- **Online augmentation:** This is used for bigger datasets, where transformations are performed on the mini-batches which are input into the classifier.

2.2 Classical data augmentation methods:

Methods listed in this section are characterized by the ease of execution. These are geometric and space transformations which are relatively simple to use but require more training time and memory. They are not always label-preserving transformations as there is a loss of information during transformations. Table 1 depicts the characteristics of various classical methods.

Table 1: Classical Augmentation Methods

S.No.	Method	Characteristics
1	FLIPPING	Horizontal and vertical flips, easiest method, non-label preserving
2	COLOR SPACE	R,G,B manipulations using color histograms; intensity modification
3	CROPPING	Centre patch cropping, random cropping, non-label preserving
4	ROTATION	Rotation by 1 to 359 degree, depending on the angle of rotation label preservation differs.
5	TRANSLATION	Spatial dimensions are manipulated
6	NOISE INJECTION	Addition of a matrix containing random values selected at random using a Gaussian distribution
7	COLOR SPACE INJECTION	Alteration of color distribution of images, RGB to Grey Scale transformation, color jittering, random color manipulation, edge enhancement and PCA.
8	KERNEL FILTERS	Use to sharpen and blur images using filters like Gaussian blur filters, vertical and horizontal edge filters.
9	MIXING IMAGES	Mixes images by averaging pixels, mixing images by averaging RGB channel values, random cropping and mixing of them to create new images
10	RANDOM ERASING	A n × m patch of an image is randomly selected and masked with either 0, 255, mean pixel values or random values; can be stacked over other techniques; non label preserving.

2.3 Modern techniques: based on deep learning

Deep learning models are extremely effective in converting high-dimensional inputs into low-dimensional representations. They transform pictures into binary classes or nx1 vectors in flattened layers. It is possible to modify neural networks' sequential processing so that transitional depictions can be distinguished from the complete data. In fully connected layers, the lower-dimensional depictions of the image data can be isolated and extracted. Table 2 describes the various deep learning methods which are used for data augmentation in brief.

Table 2: Deep Learning Augmentation Methods[6]

SNO	METHOD	Characteristics
1	Feature space augmentation	Its works using auto encoders to create new inputs, or by isolating a vector representation of a CNN, these are used as input to other machine learning algorithms; but such representations are difficult to interpret.
2	Adversarial training	Two or more networks are used such that their objectives contrast (loss function); strengthen weak spots in machine learning models, may not be useful in problems suffering from overfitting
3	GAN-based Data Augmentation	Two models are used, increases computation speed, enriches datasets, lessen the false positive rate of detection of abnormalities, and solve problems with class imbalance and overfitting.
4	Neural Style Transfer	Manipulates the sequential representation of a CNN models; transferring the style of one image to another; preserves unique content, similar to color space lighting transformations; allows transformations such as lighting variations, texture variations and artistic styles; effort demanding

3 Proposed Model: EyeGAN

3.1 Conditional GAN (CGAN)

CGAN enables the method to depend on external data for enhancing image quality. CGANs control the output of the generator at test time by giving the label for the desired image generation. In CGAN, the generator produces an image for a class using a latent space point and a label as input [7]. Image and class name are sent to the discriminator, and it determines the originality of the input. The class labels help in preventing biased image generation as GAN models tend to generate the images each operation a 4D output is obtained. The output thus obtained is the final generated output. CGANs have proven to achieve good results for image generation [8].

3.2 Proposed GAN architecture

The proposed model works similarly to CGANs, i.e., a condition is applied to the generator. In the proposed model, the class labels are sent to the generator for a particular class of fundus images generation.

3.2.1 Generator architecture

The Generator consists of three blocks, each consisting of 1 convolution layer (Leaky ReLU activated), Up-sampling, and Batch Normalization. As shown in Figure 2, the convolution layer accepts an input in the form $(32, 224, 224, 1)$ where the input is in a batch of 32 images with $(224 \times 224 \times 1)$ dimensions.

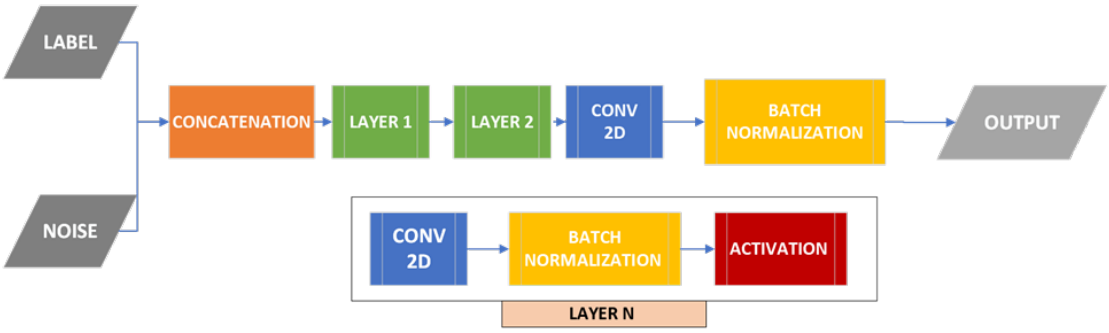


Figure 2: Generator Design of EyeGAN

At the end of each operation, a 4D output is obtained. The output thus obtained is the final generated output. To produce a single $224 \times 224 \times 1$ image, a latent vector of noise and class label are provided as inputs. Each classified input receives the class label as a parameter. A batch normalization layer and an activation layer come after each convolutional layer, except for the final one. The model employs techniques including LeakyReLU activation, a kernel of size $(5, 5)$, a stride of $(2, 2)$, and an activation function for the hyperbolic tangent in the output layer.

3.2.2 Discriminator architecture

Figure 3 shows that in the discriminator model there are two outputs and one image input $(224 \times 224 \times 1)$. The discriminator decides whether the image is real or fake and outputs the class label. Every discriminator block contains a convolutional layer, a batch normalization layer, and an activation layer. There are three such blocks which downsample the input image from $(224 \times 224 \times 196)$ to $(221 \times 221 \times 176)$ and then to $(128 \times 128 \times 160)$. The model employs techniques including LeakyReLU activation, a kernel of size $(5, 5)$, and a stride of $(2, 2)$. An activation function of sigmoid function calculates the authenticity of input, and the other output layer, which outputs the label, uses a softmax function. The complete model architecture is shown in Figure 3.

3.3 Loss Function

Based on the literature survey done, it is observed that cycle GAN is used for image generation where there is a class imbalance [9] and StyleGAN [10] generates high resolution and high diversity images. Therefore, these two are chosen for comparison with the proposed GAN. Also, to find the best loss function for the proposed model; mean squared error (MSE), mean absolute error (MAE), and binary cross-entropy are studied.

3.3.1 Binary cross entropy

The amount of the difference between the predicted probability distribution and the true probability distribution for a binary classification problem is called binary cross-entropy loss. The training data has two classes. The loss is usually

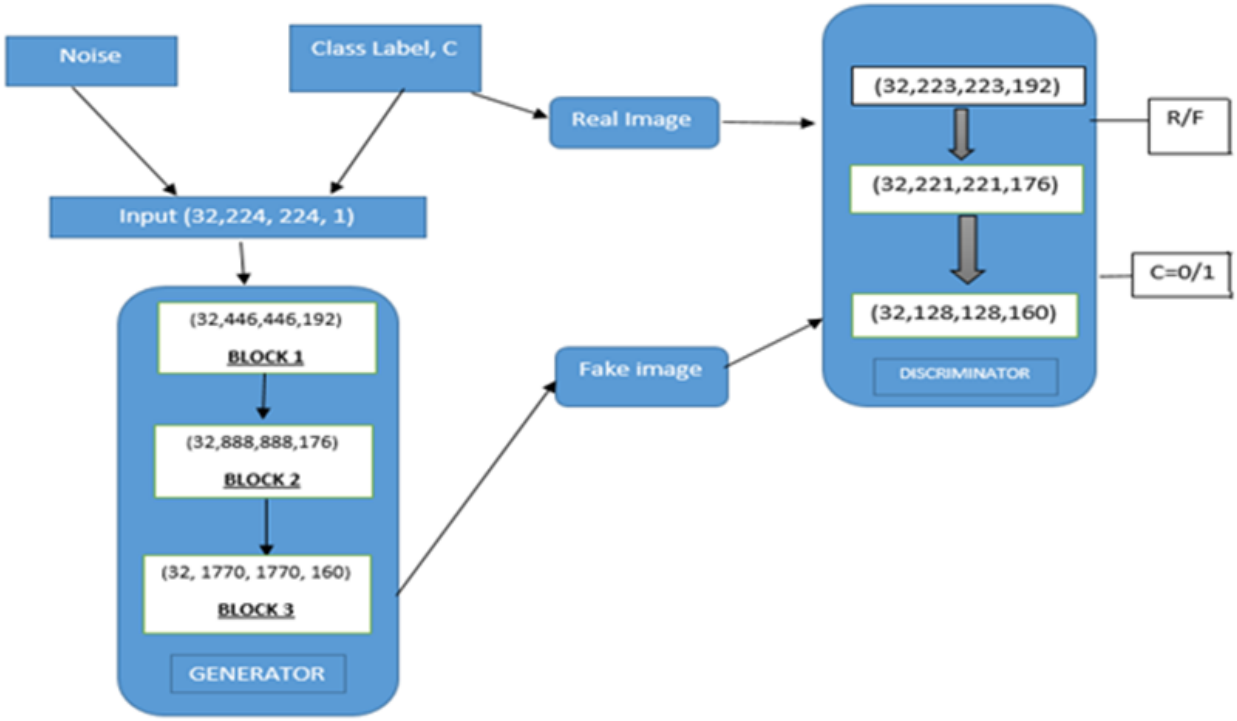


Figure 3: Overall model architecture

employed as the loss function. The binary cross-entropy loss is defined as:

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \tag{1}$$

where N is the number of rows and M is the number of columns.

3.3.2 Mean squared error (MSE)

Mean Squared Error (MSE) is a way to quantify the dissimilarity between real values and anticipated values in a regression problem. When a model is trained to produce predictions based on continuous data, this method is frequently employed as the loss function. The MSE is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

where N represents the number of data points on all variables, y is the vector of studied values for the predicted variable, and $\hat{y}$ represents the predicted values.

3.3.3 Mean absolute error (MAE)

Mean Absolute Error (MAE) analyses the average magnitude of the errors in a set of estimates while ignoring the direction. It represents the weighted average of the individual deviations between the actual value and the predictions for the test example. The MAE is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - x_i| \tag{3}$$

Table 3 shows the comparison between the three loss functions values for EyeGAN, Style GAN, and Cycle GAN. Additionally, in Figure 4, it can be seen that the binary loss function gives the least loss value. Therefore, for the proposed method, the binary loss function is used.

3.3.4 Proposed GAN Loss

- **Generator loss**: The objective is to lessen the second term in the discriminator loss equation. The generator loss function for a single created data is expressed as follows:

$$\mathcal{L}(\text{Generator}) = \min\left[\log(1 - D(G(z)))\right] \tag{4}$$

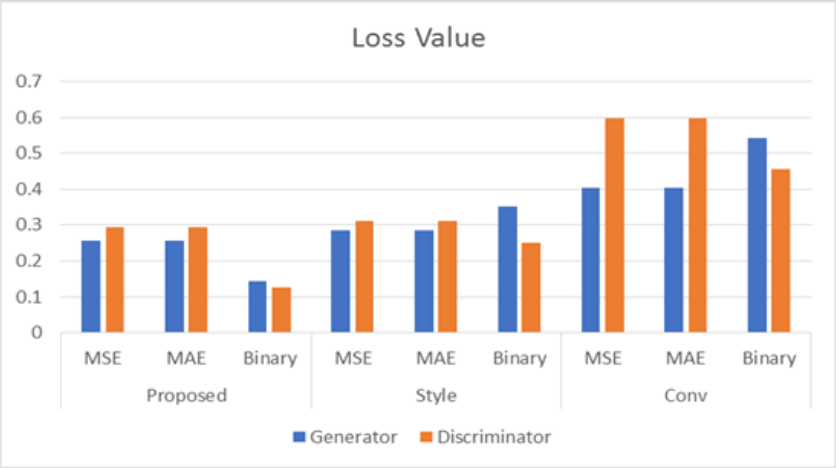


Figure 4: Comparison of Loss Functions

Table 3: Loss Value of Generator and Discriminator

GAN Model	Loss Function	Generator	Discriminator
EyeGAN	MSE	0.2575	0.2941
	MAE	0.2575	0.2941
	Binary	0.1447	0.1248
Style GAN [10]	MSE	0.2858	0.3117
	MAE	0.2858	0.3117
	Binary	0.3511	0.2512
Cycle GAN [9]	MSE	0.4026	0.5964
	MAE	0.4026	0.5964
	Binary	0.5417	0.4568

- **Discriminator loss**: There are two losses for the discriminator: one for a real input and the other for a fake input, which gives the below discriminator loss. Since the model is a conditional GAN, the loss function is represented as shown below:

$$\mathcal{L}(\text{discriminator}) = \max\left[\log(D(X)) + \log(1 - D(G(z)))\right] \tag{5}$$

$$\mathbb{E}_x[\log(D(x|y))] + \mathbb{E}_y[\log(1 - D(G(z|y)))] \tag{6}$$

The modified function has the same fundamental structure with a small modification. With this updated loss function, each part's loss is determined by factoring in the condition, y.

$$\min_G \max_D V(D,G) = \mathbb{E}_x \sim p_{\text{data}}(x)\left[\log D\left(\frac{x}{y}\right)\right] + \mathbb{E}_z \sim p_z\left[\log\left(1 - D\left(G\left(\frac{z}{y}\right)\right)\right)\right] \tag{7}$$

3.4 Database

The Messidor I dataset [11], which contains 1200 fundus color images captured with a 45-degree field of view, is one of the datasets considered. The images vary in size: 1440x960, 2240x1488, or 2304x1536. Fundus images are graded based on the number of microaneurysms, hemorrhages, and the presence of neovascularization, categorized into grades R0, R1, R2, and R3. The Messidor-2 dataset [11], an expansion of the Messidor dataset, consists of diabetic retinopathy examinations, each including two images of the eye fundus with the macula at their center (one per eye). This collection includes 1748 images, also captured with a 45-degree field of view. Neither dataset is balanced across classes. Table V shows the image distribution in the studied datasets. It is evident that class 0 has the maximum number of images, whereas in Messidor I, class 1 and class 2 have very few images, and in Messidor II, class 3 has only 75 images.

4 Experimentation and Results

4.1 4.1 Fréchet Inception Distance (FID)

Various GAN models can be judged based on the graphic quality of synthetic images. For a high-quality dataset, Fréchet Inception Distance (FID) [12] calculates the image quality and diversity, both of which are crucial. Since this is reliable and effective, this score is frequently used to assess the effectiveness of various models. This research compares the GAN model performances using this score. The FID is defined as:

$$FID = d^2((m_r, C_r), (m_g, C_g)) = \|m_r - m_g\|_2^2 + \text{Tr}\left(C_r + C_g - 2(C_r C_g)^{\frac{1}{2}}\right) \tag{8}$$

where d is the FID, (m_r, C_r) represents the real images and (m_g, C_g) represents the generated images, m and C are means and covariance, respectively. The FID score demonstrates the variations between the generated and actual images. A lower score indicates better performance. Table 4 compares the three models in terms of structure, features, and average FID score. The proposed model has fewer layers than the approaches in use, making it faster and lighter. Additionally, a DR class-wise FID score comparison is done in Table 5. The proposed model attained the least FID score for all the classes, promising the best quality image generation.

Table 4: Comparison of GAN Models

Model	Features	Average FID Score
Model 1	Feature Set 1	0.50
Model 2	Feature Set 2	0.60
Model 3	Feature Set 3	0.55

Table 5: DR Class Wise FID Score Comparison

Class	Model 1 FID Score	Model 2 FID Score
Class 0	0.50	0.60
Class 1	0.55	0.65
Class 2	0.60	0.70
Class 3	0.65	0.75

4.1.1 Data Distribution of the Studied Datasets

Figure 5 shows the original images and the respective generative images for each class for the MESSIDOR I dataset. Table 6 shows the data distribution of the studied datasets. This table also illustrates the number of generated images added to overcome the problem of data imbalance.

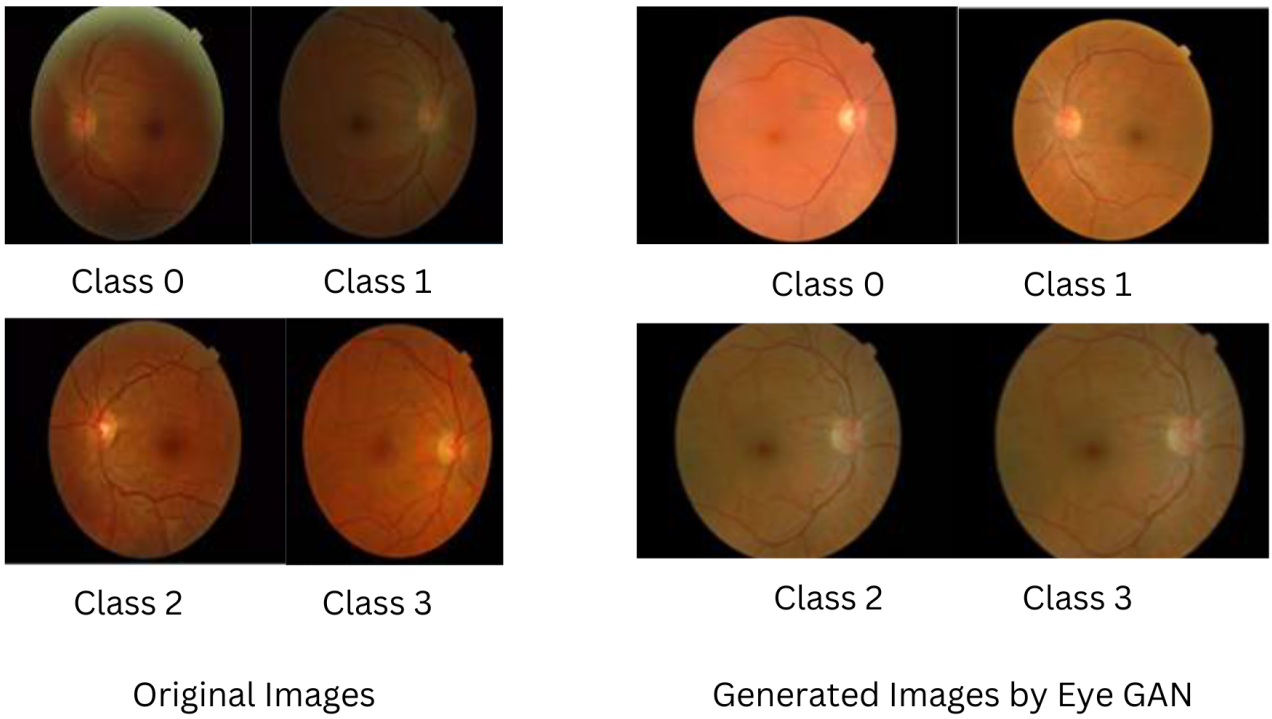


Figure 5: Comparison between original and synthetic images generated

It is observed that:

- The synthetic fundus images were noise-free.
- The EyeGAN produced real-like, good quality images for each class, thus providing an enriched dataset which can be used for improving the results by anyone.
- Both the datasets became class balanced with the application of EyeGAN.

Table 6: Data Distribution of the Studied Datasets

Dataset	Class	Number of Original Images	Images generated by EyeGAN	Total
Messidor I [10]	0	151	349	500
	1	30	470	500
	2	70	430	500
	3	149	351	500
	Total	400	1600	2000
Messidor II [10]	0	1017	0	1017
	1	270	230	500
	2	347	153	500
	3	75	425	500
	Total	1709	808	2517

5 Conclusions

With the need for a balanced and good quality dataset, attaining good accuracy of deep neural networks is conceivable. In the interest of expanding datasets and enhancing the accuracy of deep learning methods in Diabetic Retinopathy (DR) detection, a Conditional GAN-based model named EyeGAN is proposed in this study. The research is implemented on two datasets, MESSIDOR I and MESSIDOR II, which are class imbalanced datasets. By expanding the dataset, synthetic data augmentation increases its variability. EyeGAN attains the least Fréchet Inception Distance (FID) score in comparison with Cycle GAN and Style GAN. Our research demonstrates that the synthesized fundus images have qualities and properties that enrich the datasets with realistic and noise-free images. It can be concluded that the desired results are achieved. The model outperforms the existing GAN models and also balances the datasets studied. This will help in proving a strong system which will perform efficiently and overcome the problem of overfitting faced in fundus diseases detection and classification.

Declaration of Competing Interests

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Funding Declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution

Preeti Kapoor: Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Data Curation ; **Shaveta Arora**:Writing - Review & Editing, Supervision.

References

[1] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[2] H. Lee, M. Ra, and W.-Y. Kim, "Nighttime data augmentation using gan for improving blind-spot detection," *IEEE Access*, vol. 8, pp. 48049–48059, 2020.

[3] B. Liu, C. Tan, S. Li, J. He, and H. Wang, "A data augmentation method based on generative adversarial networks for grape leaf disease identification," *IEEE Access*, vol. 8, pp. 102188–102198, 2020.

[4] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020.

[5] L. Lan *et al.*, "Generative adversarial networks and its applications in biomedical informatics," *Frontiers in Public Health*, vol. 8, p. 164, 2020.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[7] Q. Jin, X. Luo, Y. Shi, and K. Kita, "Image generation method based on improved condition gan," in *2019 6th International Conference on Systems and Informatics (ICSAI)*, pp. 1290–1294, IEEE, 2019.

[8] Y. Ikeda, K. Doman, Y. Mekada, and S. Nawano, "Lesion image generation using conditional gan for metastatic liver cancer detection," *Journal of Image and Graphics*, vol. 9, no. 1, 2021.

[9] D. Li, W. Xie, B. Wang, W. Zhong, and H. Wang, "Data augmentation and layered deformable mask r-cnn-based detection of wood defects," *IEEE Access*, vol. 9, pp. 108162–108174, 2021.

[10] X. Cao, H. Wei, P. Wang, C. Zhang, S. Huang, and H. Li, "High quality coal foreign object image generation method based on stylegan-dsad," *Sensors*, vol. 23, no. 1, p. 374, 2022.

[11] P. Kapoor and S. Arora, *Applications of Deep Learning in Diabetic Retinopathy Detection and Classification: A Critical Review*, vol. 91 of *Lecture Notes on Data Engineering and Communications Technologies*, pp. 505–535. Springer Singapore, 2022.

[12] J. Yang, Z. Zhao, H. Zhang, and Y. Shi, "Data augmentation for x-ray prohibited item images using generative adversarial networks," *IEEE Access*, vol. 7, pp. 28894–28902, 2019.

Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 2 Issue 6

Article Number: 230108

Sentiment Analysis on IMDB Review Dataset

Shubham Kumar Singh*and Neetu Singla

Department of Computer Science and Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

Abstract

A computational method known as sentiment analysis is employed to ascertain the emotional undertone or attitude of a text document, such as a review, tweet, or news story. Using machine learning models, deep neural network models, and natural language processing, the method entails examining the text to determine whether it expresses positive or negative sentiment. In this study, models like Naive Bayes, Logistic Regression, LSTM, LSVM, Decision tree, and BiLSTM are utilized to conduct a sentiment analysis (SA) study on the IMDB dataset. The goal of the investigation is to evaluate how well these models perform in retrospect on movie reviews, categorizing them as positive or negative. The study investigates the effects of data pre-processing methods and hyperparameter tuning on the models' accuracy. The final results demonstrate that the BiLSTM model outperforms the other models in terms of recall, precision, and accuracy, followed by the LSTM, Logistic Regression, LSVM, Decision Tree, and Naive Bayes models. The research emphasizes the potential of deep learning models—in particular, BiLSTM in sentiment analysis tasks, as well as the significance of hyper-parameter tuning and pre-processing methods in achieving high accuracy.

Keywords: Sentiment Analysis; IMDB Review Dataset; Machine Learning Models; Data Preprocessing; Model Performance Evaluation

1 Introduction

Natural language processing (NLP) has a subfield called sentiment analysis [1, 2] that aims to recognise subjectivities, attitudes, and moods in a given textual context. It offers a new way of approaching the traditional method to classify text. Classifying text based on emotion is one of the biggest challenging areas of research in NLP, with ongoing studies in text mining. Recently, sentiment analysis based on deep learning methods such as memory networks, CNN, RNN, and BiLSTM has been extensively explored. These approaches allow for the use of multiple contexts, which can help remove characteristics from training data. However, this method needs a lot of tagged training data, which might not always be accessible. Despite advances in algorithms and methodologies, sentiment analysis still faces some unresolved issues. One limitation is the inability to classify phrases that lack overt emotional keywords, which could lead to the incorrect inference that a sentence is emotionless. Additionally, depending on the context, certain keywords may have multiple meanings, resulting in ambiguity. The system must consider these constraints and provide accurate data classification, particularly when used in sensitive applications. A Python-based programme is used in this study to analyse the IMDb dataset. The training and testing portions of the dataset are separated in a distinct way. The Naive Bayes, Logistic Regression, LSTM, LSVM, Decision Tree, and BiLSTM classifiers are developed using the training phase. The classification precision is then calculated using the testing set.

*Corresponding author: singhshubham0051@gmail.com
Received: 14 November 2023; **Revised:** 02 December 2023; **Accepted**: 04 December 2023; **Published:** 31 December 2023

DOI: 10.57159/gadl.jcmm.2.6.230108.

2 Dataset

The dataset compiled by Andrew Maas was utilized for the study, consisting of 50,000 IMDb film reviews, divided into training and testing sets with 12.5k positive and negative reviews each. The IMDb rating system was applied to categorize the reviews. Various machine learning classifiers, including Logistic Regression, Naive Bayes, LSTM, LSVM, Decision Tree, and BiLSTM, were employed on the original feature sets for text-based sentiment analysis. The balance between positive and negative reviews in the dataset is illustrated in Figure 1, which shows an equal distribution of sentiments, ensuring a fair basis for training and testing the classifiers.

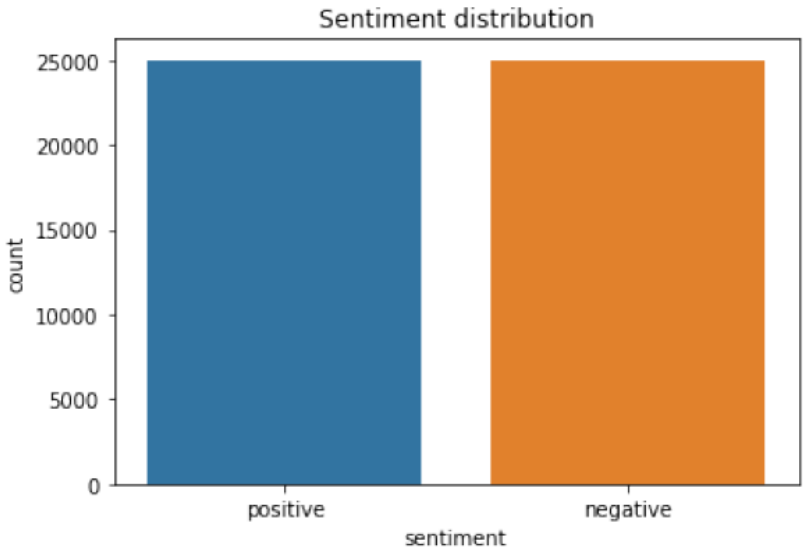


Figure 1: Distribution of Positive and Negative Sentiments in the IMDb Dataset

3 Data Division

It is common practice to split the dataset into training and testing vectors. The training vector encompasses data used for training the classifier. The training data may be divided into validation, testing, or both through various methods. Generally, training constitutes the majority of the data. In machine learning applications, an 80-20 split is often employed, where 20% of the data is used for testing and 80% for training. This division is influenced by the law of the vital few or the Pareto Principle, which is prominent in economic and financial theories. In this study, 10k reviews were selected, ensuring an equal distribution of positive and negative assessments to mitigate bias. The data was divided into 20% for testing and 80% for training. The 10-fold cross-validation method was used to eliminate bias in categorization results. Understanding the distribution of word counts in reviews is essential for preprocessing and feature engineering. Figure 2 presents histograms that illustrate the distribution of word counts in positive and negative reviews within the IMDb dataset. This information can inform decisions on text normalization and vectorization techniques.

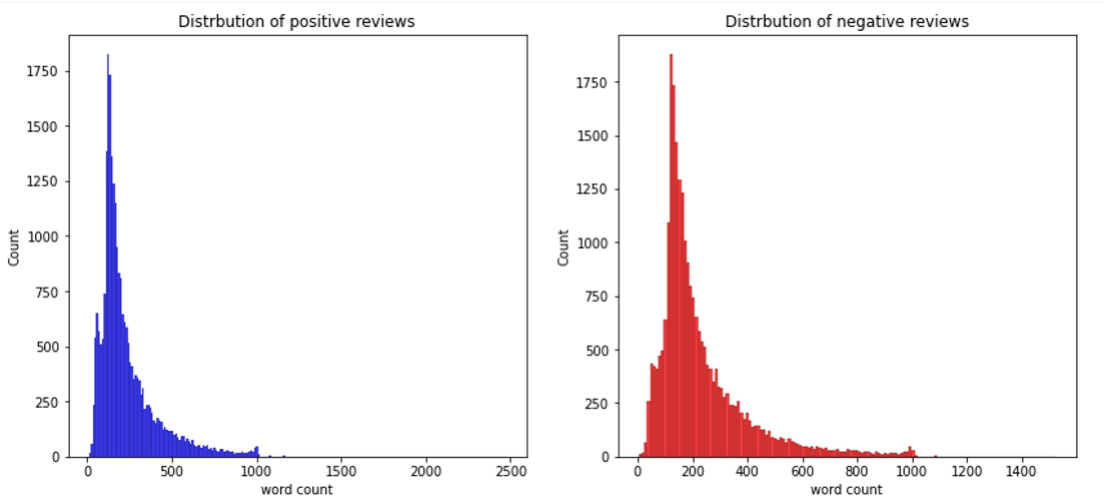


Figure 2: Histograms of Word Count Distribution in Positive and Negative IMDb Reviews

4 Related Work

In one of the previous works [3], researchers explored various methods in NLP and ML to achieve high accuracy. They specifically utilized the LSTM classifier to analyze sentiments in IMDb movie reviews, achieving a maximum classification accuracy of 89.9%. An article [4] demonstrated the use of NLP for Sentiment Analysis (SA) with conventional ML techniques. The study used Naive Bayes, Decision Tree, and Logistic Regression, with notable success in TF-IDF + Logistic Regression, yielding a validation AUC score of about 96%. Another study [5] introduced a ConvLstm neural network architecture, merging CNN and LSTM along with conditioned vector space models. This approach employed LSTM in place of the pooling layer in CNN, focusing on preserving long-term dependencies and minimizing the loss of complex local information in phrase sequences. As reported in one of the article[6], a new capsule network model named caps-BiLSTM was proposed for sentiment analysis. Integrating BiLSTM, this model showed encouraging results on various datasets, including MR, IMDB, and SST, surpassing deep learning models compared to traditional machine learning techniques. In a significant contribution [7], a semi-supervised learning technique was utilized with a limited amount of labeled data from the IMDb and YelpNYC datasets. This model outperformed baseline models like LSTM and SVM, especially in scenarios with minimal labeled data, such as when only 1% of the data is labeled. A noteworthy study [8] combined CNN and LSTM of Recurrent Neural Networks for sentiment analysis, aiming for higher accuracy with minimal loss and reduced computing time. The experiment revealed that CNN alone provided the best accuracy, while combining LSTM and CNN models yielded better efficiency in terms of speed and loss. In an insightful work [9], Sentiment Analysis on the IMDB Dataset was conducted using eight different models for classifying movie reviews. The results highlighted that the RF classifier was the most efficient, outperforming the other models in all evaluation metrics. Notably, KNN also achieved a recall comparable to RF, with a highly competitive f-measure and AUC. Yet another study detailed [10] addressed the limitations of supervised learning algorithms in sentiment analysis, mainly the need for extensive labeled data. The proposed solution was a hybrid model that combined Term Frequency-Inverse Document Frequency weighting, a lexicon for rule-based sentiment analysis, and the LSTM model. The hybrid model utilizes binary classification algorithms, such as Logistic Regression, knn, Random Forest, SVM, and Naive Bayes.

In their study [11], the authors demonstrated the use of hybrid features obtained by combining machine learning features like Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) with lexicon features such as positive-negative word count and connotation. This combination aimed to improve accuracy and complexity in classifiers like Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), and Maximum Entropy. In research [12], sentiment analysis of IMDB comments via star ratings was explored, employing SVM classification. Utilizing SVM, known for its efficacy in pattern recognition, and the TF-IDF technique, the study achieved 79% accuracy, 75% precision, and 87% recall. Notably, the research indicated that SVM outperformed logistic regression, highlighting its superiority in sentiment classification. In another study [13], researchers proposed a model combining various sentiment analysis methods to extract valuable insights and determine the optimal classifier for a specific domain, focusing on accuracy. Given the informal nature of movie reviews and their lack of strong grammatical structures, N-Grams and count vectorizer approaches were incorporated. The process involved tokenization, stemming, feature selection, and classification to transform input strings, extract word roots, select essential features, and classify movies as positive or negative. Further, in [14], diverse methods such as Naïve Bayes, SVM, Stochastic Gradient Descent (SGD), and Decision Tree were employed to achieve optimal accuracy in sentiment analysis. This study involved evaluating movie reviews, yielding 94 positive and 65 negative sentiments. The researchers found that the SVM and SGD classifiers stood out, achieving the highest accuracy at 82% and an F1 score of 81%, thus underscoring their effectiveness. Lastly, in [15], a linguistic methodology was introduced to uncover hidden genres within 600 popular science texts. By utilizing computer programs for linguistic analysis and cluster analysis, four text type clusters were identified, revealing shared linguistic traits and aiding in genre identification based on communicative purposes. An evaluation using a test set demonstrated over 70% accuracy, suggesting the method's relevance in discerning popular science genres with potential pedagogical implications.

5 Proposed Methodology

To effectively analyze the sentiment of reviews and evaluate the accuracy and processing times of various approaches, the study utilizes four advanced machine learning classification models on the IMDB dataset. Figure 3 illustrates the proposed machine learning-based model for sentiment analysis of text, which comprises six primary building blocks and several minor components functioning together as a cohesive unit. The methodology involves a procedure for conducting sentiment analysis on text data. Initially, data is collected and stored in a CSV or Excel file before being imported into the application. Preprocessing is then performed, which includes converting the entire dataset to lowercase, removing HTML tags and URLs, eliminating punctuation, and replacing chat terms and emoticons with their accurate meanings. Subsequently, tokenization is employed to cover sensitive data with recognizable identifying symbols without compromising data security. The data is then divided into training and testing sets. The dataset is subsequently used to evaluate the recall, accuracy, and precision of four machine learning models: Naive Bayes, Logistic Regression, Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM). The performance of each model is assessed in the final step to determine which model performs the best on the given dataset.

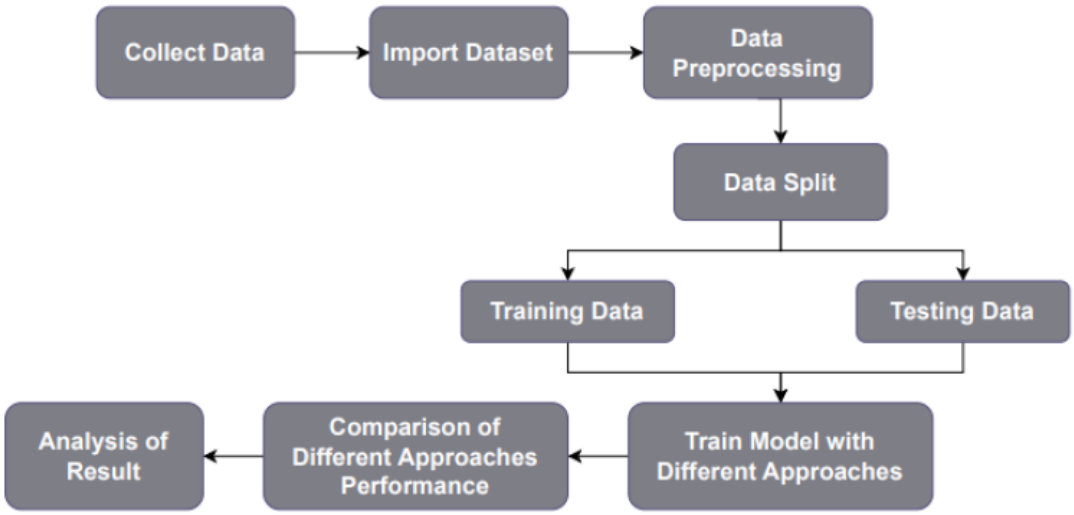


Figure 3: The proposed system flowchart for text sentiment analysis

5.1 Data Collection

The data collection procedure is essential in determining the quantity and quality of data available for sentiment analysis on the IMDb dataset. IMDb, recognized for its extensive user-generated reviews and ratings, serves as a significant online database for movies, TV series, and celebrity content. The review dataset from IMDb, which is accessible across various devices, was procured from Kaggle for this study. Kaggle is a well-known platform for data science and machine learning projects, where data scientists and machine learning experts can collaborate, exchange knowledge, and participate in diverse data science challenges. It provides users with a plethora of tools including datasets, kernels (code notebooks), competitions, and discussion forums, which are instrumental in creating, testing, and refining machine learning models.

5.2 Data Preprocessing

Data preprocessing is a vital step in sentiment analysis as it cleans and prepares the raw data for subsequent analysis. The following steps were undertaken to ready the IMDb dataset for sentiment analysis:

- **Removing HTML tags:** IMDb reviews often contain HTML tags that are extraneous to sentiment analysis. These are removed using regular expressions or the BeautifulSoup library in Python.
- **Removing special characters and punctuation:** The dataset might include special characters and punctuation that do not contribute to sentiment analysis, hence are eliminated using regular expressions or Python's string library.
- **Converting text to lowercase:** Normalizing the text to lowercase ensures that the model does not treat the same word with different cases as distinct entities.
- **Removing stop words:** Common words like "the", "and", and "is", known as stop words, are filtered out to reduce dataset dimensionality and boost model performance. Tools for this include NLTK and spaCy in Python.
- **Tokenization:** This involves segmenting text into tokens, typically single words, which is a crucial step for feature engineering, enabling the creation of a bag of words or n-grams.
- **Stemming and lemmatization:** These techniques reduce inflectional forms and sometimes derivationally related forms of a word to a common base form, thereby enhancing model performance and reducing dataset dimensionality. Lemmatization returns a word to its base or dictionary form, while stemming simply removes the suffixes of words.
- **Spell checking and correction:** To enhance the accuracy of the sentiment analysis model, it is imperative to correct spelling mistakes that might be present in the dataset. Python offers several libraries such as PySpellChecker and TextBlob that provide functionalities for spell-checking and automatic corrections.
- **Handling negation:** In sentiment analysis, correctly interpreting negation is essential since words like "not" or "never" can completely change the sentiment of a sentence. Techniques to address negation include using a separate feature to indicate negation or appending a negation prefix to the affected words.

- **Removing rare and frequent words:** Words that are too rare or too frequent may not contribute significantly to sentiment analysis. Eliminating these words can reduce the dimensionality of the dataset and enhance model performance. Methods for this include setting frequency thresholds, removing stopwords, applying lemmatization or stemming to consolidate variations of the same word, and maintaining curated lists of terms for removal, especially noisy or domain-specific ones. These strategies can significantly refine the dataset, thus improving the data quality for sentiment analysis projects.

Figure 4 displays the most frequent words in both positive and negative reviews after the data preprocessing steps. This analysis helps in understanding the distribution of words in the dataset. Figure 5 illustrates the most commonly used words in the preprocessed dataset. This analysis provides insights into the prevalent terms within the dataset, which can be essential for understanding sentiment trends.

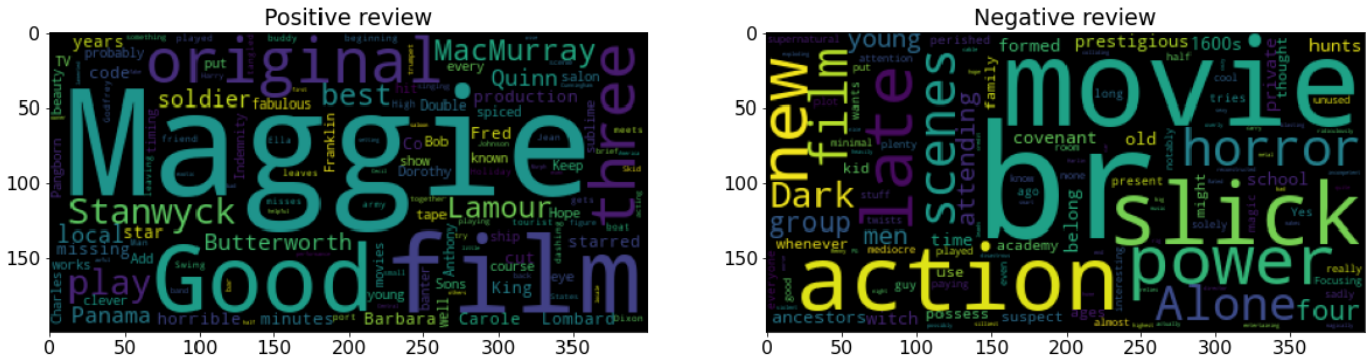


Figure 4: Most Frequent Words in Positive and Negative Reviews

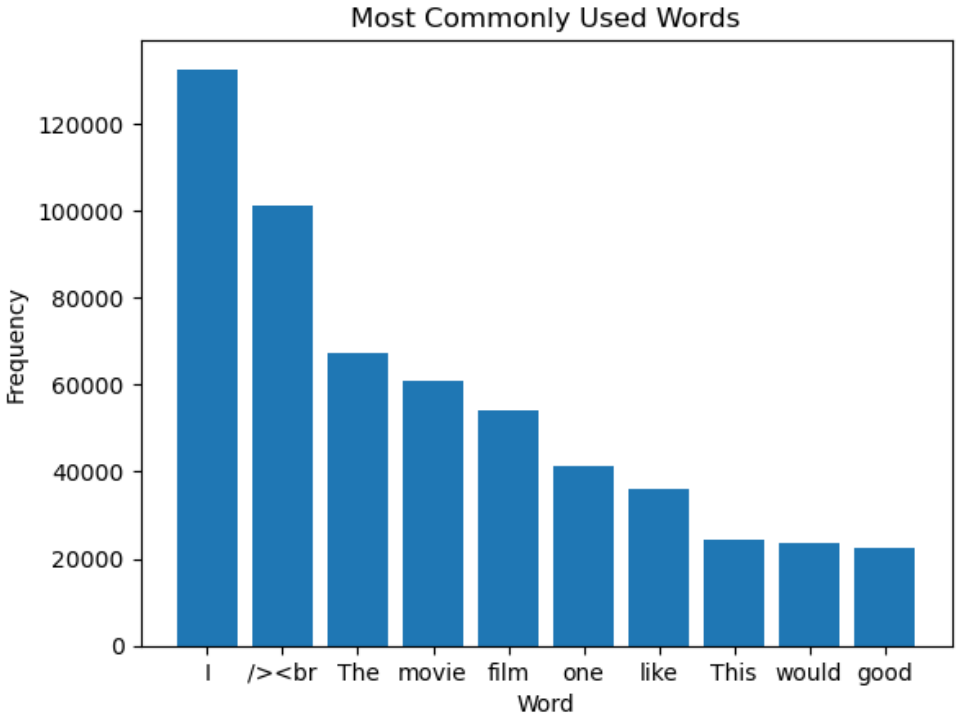


Figure 5: Most Commonly Used Words

6 Feature Engineering

Feature engineering is essential for converting unstructured textual data into a format that machine learning models can understand and utilize efficiently in the fields of NLP, which stands for natural language processing, and sentiment analysis. The *Bag of Words* (also referred to as BoW) and *Term Frequency-Inverse Document Frequency* (also termed as TF-IDF) are basic methods used for feature engineering in my project, and they are covered in this section.

- **Bag of Words (BoW)**: The BoW model treats documents as collections of individual words, focusing on the presence and frequency of terms while ignoring the order. In this study, BoW involves tokenizing text into words to create a vocabulary of unique terms and then representing each document as a vector, with values indicating the frequency of each word. Although this results in a high-dimensional matrix, dimensionality reduction techniques like Principal Component Analysis (PCA) can be applied to make the computational complexity manageable. By converting text to numerical data, BoW facilitates the understanding and analysis of text by machine learning models.
- **TF-IDF**: Term Frequency-Inverse Document Frequency (TF-IDF) is a sophisticated method that evaluates the importance of a word in a document relative to a collection of documents, the corpus. Term Frequency (TF) emphasizes words that occur more frequently in a document, whereas Inverse Document Frequency (IDF) gives more weight to words that are rare across the corpus. In this study, TF-IDF is utilized to extract significant text features, thus enhancing the predictive capabilities of the model. It involves calculating the TF for each word, determining the IDF to weigh word scarcity, and then generating vectors for documents based on these TF-IDF scores. This method is invaluable for identifying and stressing the unique and relevant terms in each document, proving to be an essential tool for sentiment analysis.

6.1 Model Architecture

The architecture of a model encompasses several integral stages such as model selection, training, and evaluation, each of which plays a crucial role in the overall machine learning process.

- **Model Selection**: The initial step involves selecting appropriate machine learning models for sentiment analysis. A variety of models are considered, including neural networks, Support Vector Machines (SVMs), logistic regression, and Naive Bayes. Each of these models is adept at extracting meaningful patterns from data, managing high-dimensional spaces, and recognizing subtle sentiment indicators within text. The deliberate choice of the most fitting model is essential for accurate sentiment classification and is a key component of the methodology.
- **Model Training**: After selecting a model, it is trained using the preprocessed IMDb review dataset. Training involves presenting the model with the data and fine-tuning its parameters to optimize performance for the task at hand.
- **Model Evaluation**: After training, the model undergoes an evaluation phase to determine its effectiveness in sentiment analysis on the IMDb dataset. The dataset is divided into training and testing sets, with the former used for developing the model and the latter for gauging its performance. Evaluation metrics such as accuracy, precision, and recall are used to assess the model's capabilities.

The results from the evaluation guide further refinements of the model. If necessary, additional rounds of training and evaluation are conducted to achieve the targeted performance metrics.

7 Model

7.1 Naive Bayes

The Naive Bayes classifier is based on applying probabilities to text sentiment analysis, utilizing the principles of Bayes' theorem. The theorem posits that the probability of a hypothesis, in this context, a sentiment label, given observable evidence, which is the textual data, is proportional to the probability of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis.

$$P(M|N) = \frac{P(N|M)P(M)}{P(N)} \tag{1}$$

In sentiment analysis, Naive Bayes assesses the probability of a text belonging to certain sentiment categories (e.g., positive or negative) based on the presence of particular words or text features. The approach is termed "Naive" because it assumes that all features (or words) in the text are independent of each other. This simplification expedites the computation process and increases the efficiency of the algorithm, albeit at the expense of possibly ignoring actual interdependencies between features. To prepare Naive Bayes for sentiment analysis, it is trained on a dataset where texts are pre-labeled with sentiment classes. Once trained, the algorithm is capable of predicting sentiments for new, unlabeled texts. Despite its simplicity, Naive Bayes has demonstrated effectiveness in sentiment analysis, particularly when working with large datasets and straightforward feature models. The performance of the Naive Bayes model is visually represented in the confusion matrix shown in Figure 6. This matrix provides insights into the model's precision and recall for sentiment classification.

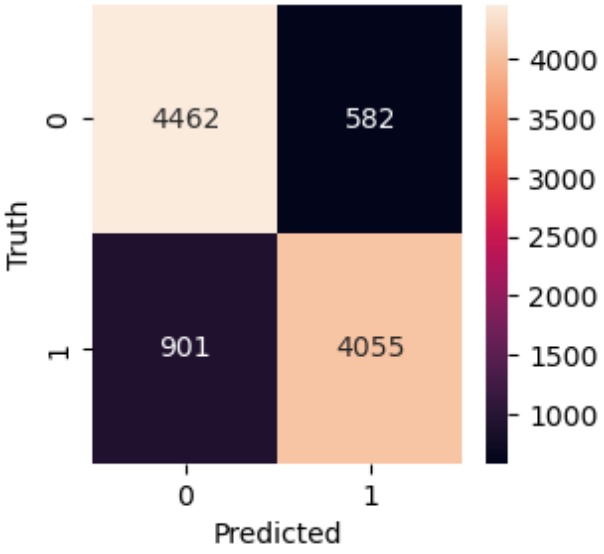


Figure 6: Confusion Matrix for Naive Bayes Model

7.2 Logistic Regression

Logistic regression is a statistical method frequently employed in sentiment analysis of textual data. It uses the logistic function to estimate the probability that a given piece of text belongs to a certain sentiment class, such as positive or negative, based on its constituent words or features. In the context of sentiment analysis, logistic regression applies a logistic function to model the probability of text belonging to a specific sentiment class. This is based on the linear combination of the text's features, with optimal feature weights determined through the process of maximum likelihood estimation on labeled data.

After the training phase, the algorithm can then be used to predict sentiment classes for new, unlabeled texts. Its effectiveness in sentiment analysis is well-established, with applications ranging from binary to multiclass classification tasks. The logistic regression model's performance is further elucidated by the confusion matrix depicted in Figure 7, which outlines the number of true positive, false positive, true negative, and false negative predictions made by the model. This visual representation aids in comprehending the precision and recall of the model. The logistic regression model is described mathematically as:

$$P(X) = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_n x_n)}} \tag{2}$$

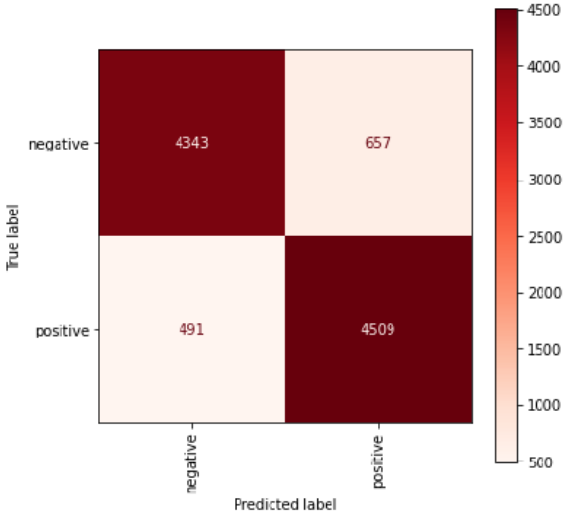


Figure 7: Logistic Regression Confusion Matrix

where $P(X = 1|z)$ is the probability of the binary outcome variable being 1 (for instance, a positive sentiment) given the predictor variables z. The $\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n$ are the coefficients or weights corresponding to the predictor variables $x_1, x_2, \ldots, x_n$.

The logistic or sigmoid function is employed to convert the linear combination of predictor variables and their coefficients into a probability score that ranges between 0 and 1. This relationship is also expressed in the form of log-odds or the logit function:

$$\log\left(\frac{P(X=1|z)}{1-P(X=1|z)}\right) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_n x_n \tag{3}$$

This equation is utilized to calculate the weights or coefficients of the predictor variables using maximum likelihood estimation from a labeled dataset. The left side of the equation, the log odds, indicates the likelihood that the binary outcome variable is equal to 1.

7.3 Long Short-Term Memory

An RNN model architecture that works well for evaluating sentiment in textual data is LSTM. Its goal is to deal with the problem of vanishing gradients, which may happen in conventional RNNs and make it impossible to maintain long-term dependencies in the input sequence. Specialized memory cells that may selectively preserve or delete data at various time steps are incorporated into LSTM in order to address this issue. The flow of information is controlled by various "gates", including an input gate, a forget gate, and an output gate. Figure 8 illustrates these components of an LSTM unit. Three components make up an LSTM: the input gate, forget gate, and output gate, which control the flow of data

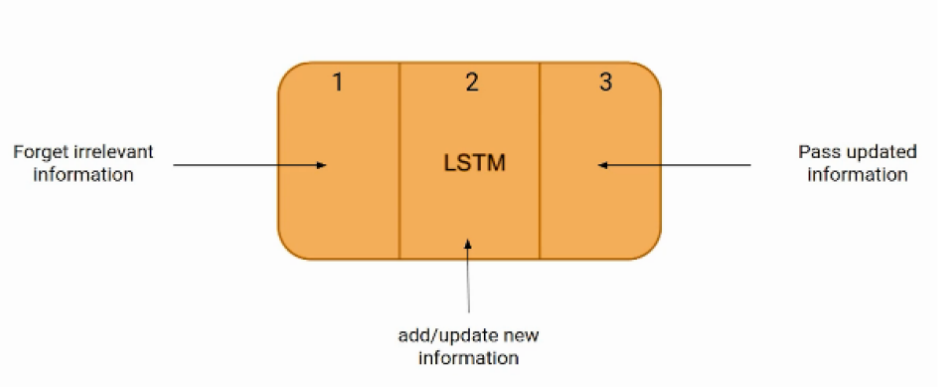


Figure 8: Illustration of LSTM Unit with Input, Forget, and Output Gates

into and out of the memory cell. These components are responsible for managing the flow of data within the LSTM. The extent to which fresh data is permitted to enter the cell is governed by the input gate, while the forget gate is responsible for determining the level at which previous information is eliminated from the cell. The amount of the cell's current state that is used to create predictions is controlled by the output gate. The LSTM's ability to forecast sentiment on new, unlabeled text data post-training can be assessed by its confusion matrix, as depicted in Figure 9. LSTMs can be trained

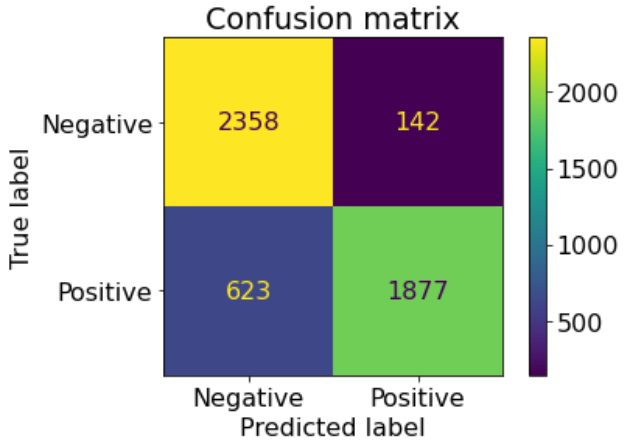


Figure 9: Confusion Matrix for LSTM Model Sentiment Classification

on labelled datasets of text data and are frequently used in sentiment analysis tasks. The network can be used to forecast the sentiment of fresh, unlabeled texts after training. For sentiment analysis tasks involving longer input sequences and more complicated features, LSTMs are particularly effective.

7.4 Bidirectional Long Short-Term Memory

A neural network architecture known as a BiLSTM is useful for analysing the sentiment of textual data. It is an expansion of the LSTM architecture that considers the forward and backward information flow of the input sequence. A BiLSTM is made up of two independent LSTM layers, one for the forward direction of the sequence of inputs and another for the backward direction. The output of each LSTM layer is then combined to generate a single output containing data from both the forward and backward directions. The ability of the BiLSTM architecture to collect contextual data from both the forward and backward directions of the input sequence can increase the accuracy of sentiment analysis. To train BiLSTMs, labelled datasets of textual data that pair each text with a corresponding sentiment class label are used. The network can be used to forecast the sentiment of fresh, unlabeled texts after training. For sentiment analysis tasks requiring longer input sequences and more complicated feature sets, BiLSTMs perform particularly well. The BiLSTM architecture, which accounts for both forward and backward information flow in the input sequence, is represented in Figure 10.

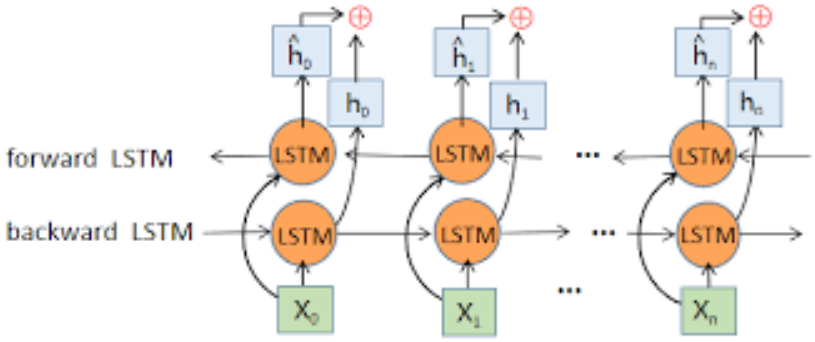


Figure 10: Bidirectional LSTM (BiLSTM) Network Architecture

7.5 Linear Support Vector Machine

For binary classification tasks, LSVM is intended. Finding the hyperplane in the feature space that best separates the two classes is its main objective. When used for sentiment analysis on textual data, LSVM can tell whether a text carries a positive or negative sentiment.

Prior to using LSVM for sentiment analysis, text data must be transformed into numerical features. A bag-of-words model, where each text document is represented as a vector of word frequencies or binary indicators, is one way to accomplish this. After that, we can use a labelled dataset of text documents to train the LSVM. In this dataset, each text document has a binary label that indicates whether it carries positive or negative sentiment. The maximum distance of the margin from the nearest data points in each class determines the decision boundary that the LSVM creates in the training phase, which improves the separation between the two classes. The sentiment of new and unlabeled text data can be predicted using the decision boundary once it has been established. Due to its ease of implementation and high accuracy on many text classification tasks, LSVM has gained popularity as an algorithm for sentiment analysis.

$$y = s^{\top} p + m \tag{4}$$

where y is the predicted sentiment label for a given text document, p is the feature vector representing the text data, s is the weight vector learned during training, and m is the bias term. The LSVM gains knowledge of the ideal s and m values during training that maximise the margin between the two classes. The hyperplane formed by the bias term m and weight vector m then establishes the decision boundary. The work undertaken merely apply the learned decision boundary to the feature vector x to produce the predicted label y in order to predict the sentiment of fresh and unlabeled text data. The text is categorised as having a positive sentiment if y is positive, and as having a negative sentiment if y is negative. The LSVM model works by finding the hyperplane that best separates two classes within the feature space, as visualized in Figure 11.

7.6 Decision Tree

A decision tree is an algorithm used to classify data, especially suited for classification tasks. It achieves this by recursively dividing the feature space into subsets based on the input feature values and assigning a class label to each leaf node in the resulting tree structure. Decision Trees are renowned for their versatility and interpretability in the realm of machine learning, excelling at deciphering complex decision-making processes by breaking them down into a sequence of clear, rule-based choices. The methodology involves the meticulous selection of relevant features from the dataset, which primarily comprises movie reviews.

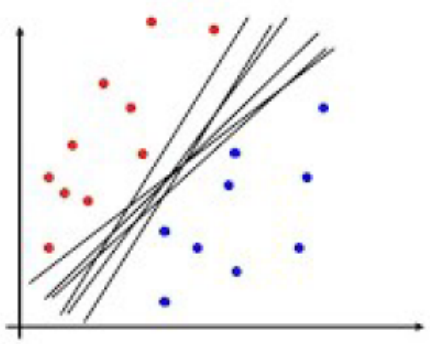


Figure 11: Illustration of LSVM Decision Boundaries

These features include words, phrases, and various linguistic attributes, forming the basis of the analysis. The Decision Tree algorithm orchestrates the construction of a hierarchical tree structure, where each internal node represents a critical decision point based on a specific feature, and each leaf node encapsulates a sentiment label, distinguishing between positive and negative sentiments. When a new movie review is analyzed, it passes through the Decision Tree, starting from the root and moving toward a leaf node. The trajectory is determined by the presence or absence of certain features within the review. At each internal node, a decision is made as to which branch to follow, guided by the content of the review. Ultimately, as the review arrives at a leaf node, the model assigns a sentiment label. If a leaf node is labeled "positive," the review is categorized as having a positive sentiment; conversely, if the path leads to a leaf labeled "negative," the review is deemed to express a negative sentiment. Within the broader context of the study, the Decision Tree Model plays a crucial role in achieving the objective of sentiment analysis. Its transparency and ability to process text data enable it to elucidate the reasoning behind particular sentiment classifications. Moreover, it complements other models used in the research, enhancing the understanding of the effectiveness of decision tree-based approaches in sentiment classification.

8 Results and Discussion

The sentiment analysis project on the IMDb dataset utilized Python for implementation. The dataset was divided into two sets, with 20% reserved for testing and the remaining 80% used for training the models. A diverse array of methodologies was employed to classify the sentiment of movie reviews effectively. These included the application of directional LSTM (Long Short-Term Memory) and Bidirectional LSTM models, which are known for their ability to capture intricate text patterns. Moreover, the predictive capabilities of Linear Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and Decision Tree algorithms were harnessed. This comprehensive approach enabled a rigorous evaluation of the performance of these models in identifying positive and negative sentiments within movie reviews, contributing valuable insights to the field of sentiment analysis.

Table 1: Performance of Classification Model

Name	Type	Performance		
		Accuracy	Precision	Recall
Model 1	Naive Bayes	0.86	0.86	0.87
Model 2	Logistic Regression	0.89	0.90	0.88
Model 3	LSTM	0.91	0.89	0.92
Model 4	BiLSTM	0.91	0.89	0.94
Model 5	Linear SVM	0.89	0.90	0.88
Model 6	Decision Tree	0.70	0.70	0.71

From the data presented in the table, it is concluded that the BiLSTM model surpasses the others in terms of accuracy, recall, and precision. The training process for Model A is visualized in Figure 12, showcasing the trends in accuracy and loss across epochs for both training and validation datasets. Similarly, the training and validation progress for Model B can be observed in Figure 13, which indicates the model's learning curve and generalization capability over successive epochs.

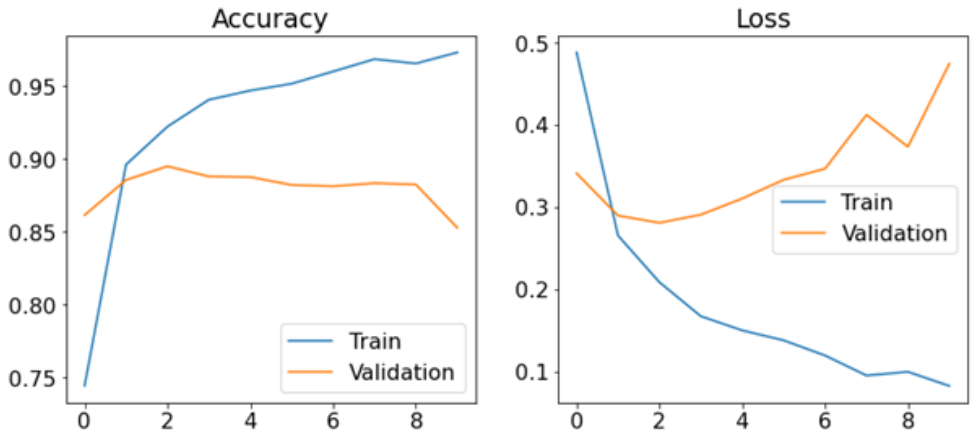


Figure 12: Training and Validation Accuracy and Loss Over Epochs for Model A

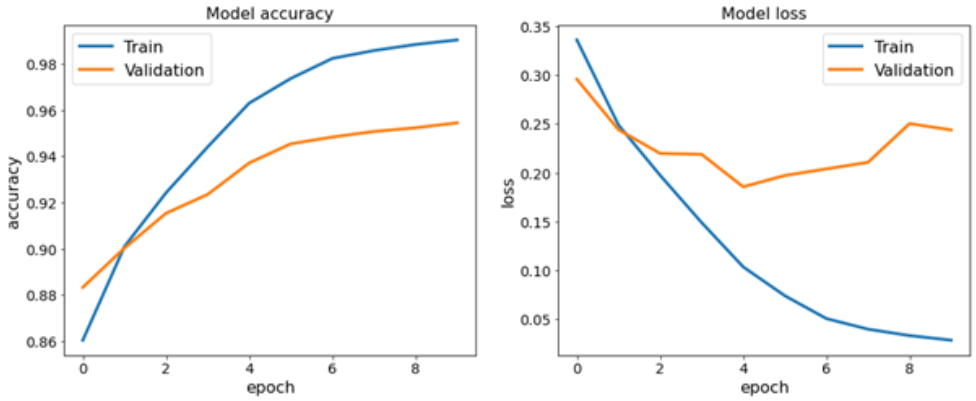


Figure 13: Training and Validation Accuracy and Loss Over Epochs for Model B

9 Conclusion

The study on sentiment analysis of the IMDB review dataset has successfully demonstrated the capabilities of various machine learning models, highlighting the superior performance of the BiLSTM model. The research revealed that deep learning models, particularly BiLSTM, offer significant advantages in processing and analyzing natural language data. These models not only excel in accuracy but also provide insights into the nuances of sentiment analysis in large datasets. Despite the promising results, the study acknowledges limitations in dataset diversity and model generalizability. It suggests that future research could focus on expanding the dataset variety and exploring more complex model architectures to enhance the accuracy and applicability of sentiment analysis techniques. This research contributes to the understanding of machine learning applications in natural language processing, particularly in sentiment analysis. It lays a foundation for future studies to build upon, potentially leading to more robust and versatile sentiment analysis tools for diverse datasets.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

Neetu Singla: Methodology, Software, Validation, Investigation, Data curation, Original draft writing, visualization; **Shubham Kumar Singh**: Conceptualization, Formal analysis, Resource management, Review and editing, Supervision.

References

[1] I. Chaturvedi *et al.*, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," *Information Fusion*, vol. 44, pp. 65–77, 2018.

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[3] S. M. Qaisar, "Sentiment analysis of imdb movie reviews using long short-term memory," in *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, IEEE, 2020.

[4] S. Tripathi *et al.*, "Analyzing sentiment using imdb dataset," in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, 2020.

[5] A. Hassan and A. Mahmood, "Deep learning approach for sentiment analysis of short texts," in *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*, IEEE, 2017.

[6] Y. Dong, Y. Fu, L. Wang, Y. Chen, Y. Dong, and J. Li, "A sentiment analysis method of capsule network based on bilstm," *IEEE Access*, vol. 8, pp. 37014–37020, 2020.

[7] Y. Pan, Z. Chen, Y. Suzuki, F. Fukumoto, and H. Nishizaki, "Sentiment analysis using semi-supervised learning with few labeled data," in *2020 International Conference on Cyberworlds (CW)*, IEEE, September 2020.

[8] S. Mathapati *et al.*, "Collaborative deep learning techniques for sentiment analysis on imdb dataset," in *2018 Tenth International Conference on Advanced Computing (ICoAC)*, IEEE, 2018.

[9] M. Yasen and S. Tedmori, "Movies reviews sentiment analysis and classification," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, IEEE, 2019.

[10] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "Lstm, vader and tf-idf based hybrid sentiment analysis model," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

[11] B. Harish, K. Kumar, and H. Darshan, "Sentiment analysis on imdb movie reviews using hybrid feature extraction method," 2019.

[12] N. G. Ramadhan and T. I. Ramadhan, "Analysis sentiment based on imdb aspects from movie reviews using svm," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 7, no. 1, pp. 39–45, 2022.

[13] A. Singh, C. Kulkarni, and N. A. Ayan, "Sentiment analysis of imdb movie reviews," 2022.

[14] P. Gunawan, T. Alhafidh, and B. Wahyudi, "The sentiment analysis of spider-man: No way home film based on imdb reviews," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 6, no. 1, pp. 177–182, 2022.

[15] A. Lieungnapar, R. Todd, and W. Trakulkasemsuk, "Genre induction from a linguistic approach," *Indonesian Journal of Applied Linguistics*, vol. 6, no. 2, pp. 319–329, 2017.

Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 2 Issue 6

Article Number: 230109

Breast Cancer Detection using Machine Learning Algorithms

Bhoomi Jain*and Neetu Singla

Department of Computer Science and Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

Abstract

Machine learning employs classification methods on datasets. The Machine Learning repository provided the cancer datasets that were used in this study, which were used for categorization. Breast cancer databases come in two varieties. There are various numbers of characteristics dispersed among these datasets. Breast cancer observes around 14% of all female cancers. One in every 28 women will develop breast cancer. To analyse patterns in datasets, machine learning algorithms like SVM, KNN, and decision trees are used. Computers are able to "learn" from their past mistakes and come up with solutions that are difficult for humans to come up with. According to the study, there are many effective algorithms for analysing the properties of data sets. This study compares and implements several well known classification methods, including Decision Trees, K Nearest Neighbor, SVM, Bayesian Network, and Naive Bayes on the Wisconsin Diagnostic dataset by calculating its classification accuracy, and its sensitivity and specificity value.

Keywords: Breast Cancer Detection; Machine Learning Algorithms; Wisconsin Diagnostic Dataset; Algorithm Performance Comparison; SVM and Decision Trees

1 Introduction

As time progresses, there are increasingly more individuals who might develop cancer due to various causes. The Kaggle website hosts an extensive array of cancer data for research purposes. Data are available in multiple formats, such as text, image, micro-array, gene expression, and others. Cancer data can be categorized into two types: Malignant (M) and Benign (B) [1]. Benign tumors are considered less harmful as their cells do not proliferate, whereas malignant tumors are harmful and carcinogenic once they start growing inside a human body. The dataset used in this study is a large dataset with unstructured data sourced from the UCI ML repository, related to breast cancer. Initially, the data undergoes pre-processing using machine learning (ML) techniques, followed by data cleaning, data selection, determining variable dependencies, and removing independent variables. A breast cancer dataset is labeled and classified as malignant or benign using ML algorithms [2]. This paper comprises six sections. The introduction is addressed in the first section [3]. The second section reviews the literature from esteemed authors in this domain. The third section outlines the methodology and machine learning techniques applied to this dataset. The fourth section discusses the procedures for data acquisition and calculations to determine accuracy. The fifth section presents the experiment conducted for data analysis, using line chart graphs to illustrate the results [4]. The sixth and final section of the paper, followed by references, presents the conclusion and future research directions. A machine learning experiment is more likely to be successful if it is well-planned, executed, and the results are rigorously evaluated [5].

*Corresponding author: bhoomij1230@gmail.com
Received: 14 November 2023; **Revised:** 01 December 2023; **Accepted**: 04 December 2023; **Published:** 31 December 2023

DOI: 10.57159/gadl.jcmm.2.6.230109.

2 Related Work

This section discusses previous research on machine learning techniques utilized by researchers for diagnosing breast cancer. Arpita Joshi and Ashish Mehta compared the classification results obtained using Random Forest, KNN, SVM, and Decision Tree methods. The Wisconsin dataset from the UCI repository was employed in their study.

Table 1: Summary of Research Papers on Breast Cancer Diagnosis Using Machine Learning Techniques

Algorithms	Datasets	Results
Naïve Bayes, SVM, J48, GRNN	Breast Cancer	GRNN & J48: 91%, Naïve Bayes & SVM: 89%
Decision Tree	WBC, WDBC, Breast Cancer	Feature selection improves WBC: 97%, Breast Cancer: 71.45%
SVM, C4.5, Naïve Bayes, KNN	WBC	SVM outperforms others: 97.13%
Ensemble, Naïve Bayes, SVM	WDBC	Ensemble & NB: 97%, SVM: 98.5%
Naïve Bayes, J48	WDBC	Naïve Bayes: 98%
MLP, J48, Rough Set	Breast Cancer	J48: 80%, MLP: 76%, Rough Set: 72.3%
IBK, SMO, BF Tree	WBC	SMO: 96.2%, IBK: 95.9%, BF Tree: 95.5%
J48, SMO, MLP, Naïve Bayes, IBK	WBC, WDBC, WPBC	WBC: J48 & MLP: 97.5%, WDBC: SMO: 98%
Classification: KNN, SVM, Naïve Bayes, K-means	WPBC	SVM & C5.0: 82%
Naïve Bayes, C4.5, SVM	WPBM	Naïve Bayes: 68%, C4.5: 74%, SVM: 75.75%

The most effective classifier, according to the simulation results, was KNN, followed by Random Forest, SVM, and Decision Tree. By integrating these techniques with feature selection/extraction methods, David A. Omondiagbe, Shanmugam Veeramani, and Amandeep S. Sidhu evaluated the performance of SVM, ANN, and Naive Bayes using the WDBC Dataset [2]. SVM-LDA was chosen over other methods due to its longer computation time, as indicated by the simulation results. Furthermore, data mining is frequently utilized in the medical field to predict and classify rare events, thereby aiding in the understanding of incurable diseases like cancer. The classification outcomes of data mining offer hope for early detection of breast cancer, which is why it is applied in this study. A summary of various research papers on breast cancer diagnosis using machine learning techniques is presented in Table 1.

3 Methodology

The methodology section outlines the application of various machine learning classifiers to the dataset. Specifically, Support Vector Machine (SVM), Decision Tree (C4.5), Logistic Regression, K-Nearest Neighbors (KNN), and Random Forests were utilized to analyze the data. The primary objective was to identify the most efficient and reliable algorithm for breast cancer detection [4].

3.1 Dataset

The dataset for the experiments as shown in Table 2 was sourced from Kaggle, focusing on the Wisconsin dataset which details characteristics of affected cell structures in breast cancer. This includes parameters like cell thickness, uniformity in cell size and shape, bare nuclei, single epithelial cell size, bland chromatin, normal nucleoli, and mitosis [1]. The dataset comprises 7,858 cases, organized into four expanded folders. Each folder corresponds to two types of tumors: benign and malignant.

Table 2: Description of the Wisconsin Diagnosis Breast Cancer (WDBC) Dataset

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Diagnosis Breast Cancer (WDBC)	32	569	2

3.2 Machine Learning (ML) Techniques for Classification

Several machine learning techniques are employed for data classification. These include Multilayer Perceptron, Bayesian Network, Naive Bayes, SVM, Decision Tree, Random Forest, and KNN. Key features of each method are as follows:

1. Support Vector Machine (SVM): SVM is is recognized as one of the effective approaches in the realm of machine learning, particularly when implementing kernel functions.

Its applications are diverse, encompassing facial recognition, database marketing, recommendation systems, text categorization, and cancer prediction, among other domains.

2. Random Forest (RF): Random Forest (RF) is an ensemble classifier that is based on the Decision Tree algorithm. RF is known to process large datasets, but it operates at a slower pace compared to other classifiers. RF generates a multitude of classification trees without the need for pruning. The pruning strategy, commonly associated with Classification and Regression Trees (CART), reduces the size of the tree by splitting the data into two subsets to find the best predictor in subsequent iterations. RF is capable of processing datasets with missing values and can estimate those missing values [6].

3. K-Nearest Neighbours (KNN): A classifier that uses the distance measure is called k-Nearest Neighbour. It is known as lazy learning or instance-based learning. The closest instance is used to complete the task locally [5]. The Manhattan distance method or the Euclidean distance are used to measure the distance. The classification in this technique is done using the smallest distance that was measured. The cost of learning the model is quite low, but it depends on the number of examples; as the number of instances rises, the cost climbs as well.

4. Bayesian Network: The probabilistic link between the relevant variables is represented by a directed acyclic graph called a Bayesian Network [2]. Each node represents a random (stochastic) variable with two or more potential states. It uses a set of variables on other variables to numerically deduce the probabilistic outcomes. It also has a reputation as a Belief Network (or Causal Probabilistic Network).

5. Decision Tree: A binary tree is constructed using the features present in datasets using the decision tree, a potent classification algorithm. ID3, C4.5, C5, J48, and CART are examples of popular algorithms. The process for selecting the root node is quite important. To find the variable and build a decision-making tree, this decision tree makes use of mathematical techniques like the Gini index, entropy, information gain, the chi-square test, etc. By dividing the variable into subgroups, homogeneity order must be preserved [7].

6. Naive Bayes: Conditional probability is the foundation of this classifier. The attributes included in the dataset are thought to be independent and reliable. In order to make it more effective, fewer parameters are used. This classifier can be used for applications such as sentiment analysis, language detection, and spam detection [8].

The comparative analysis of KNN, Naïve Bayes, and Random Forest in terms of various parameters is summarized in Table 3.

Table 3: Comparison Among KNN, Naïve Bayes, and Random Forest

Parameter	KNN	Naïve Bayes	Random Forest
Time Complexity (Training Phase)	O(1)	O(Nd)	O(MKlogN)
Problem Type	Classification and Regression	Classification	Classification and Regression
Accuracy	High	Requires a large number of records for high accuracy	High
Model Parameter	Non-Parametric	Parametric/Non-Parametric	Non-Parametric

4 Experimental Environment

4.1 Dataset Acquisition

The Breast Cancer Wisconsin Diagnostic dataset, obtained from the University of Wisconsin Hospitals Database, was utilized for the analysis[1]. This dataset provides comprehensive insights into the characteristics of breast cancer cases. It encompasses a total of 569 instances of Breast Cancer Wisconsin, with a distribution of 212 malignant (37.26%) and 357 benign (62.74%) cases, classified into two categories: malignant and benign.

4.2 Preprocessing

The initial data samples are acquired with a variety of attributes and values, often containing a wide range of issues such as outliers, noisy data, duplicates, missing values, and skewed data. To address these issues, preprocessing of the data is necessary. The data cleaning process involves eliminating or reducing missing data and noisy information. This can be achieved by deleting tuples, inputting missing values, and replacing numerical values with the mean attribute or the attribute mean of the corresponding class.

Additionally, data preprocessing techniques like feature selection, dimension reduction, and feature extraction are employed to modify data collection, making it compatible with machine learning algorithms [8].

4.3 Feature Extraction

After preprocessing, feature extraction is the subsequent step, where relevant features significant for breast cancer detection are identified and extracted from the pre-processed images. Techniques for feature extraction may include edge detection, texture analysis, or shape analysis. Following feature extraction, feature selection methods are employed to choose the most pertinent features that could enhance the machine learning model's accuracy. Some common feature selection methods are mutual information, principal component analysis (PCA), and recursive feature elimination [9].

4.4 Model Training and Validation

Post data preprocessing, machine learning methodologies such as classification, prediction, and estimation are applied to develop the model. To prevent overfitting, the model is trained and validated on a dataset separate from the one used for training. Test datasets estimate model error, while training sets are used for model construction. Techniques like Artificial Neural Networks (ANN), Decision Trees, SVM, and Bayesian Networks are utilized for predicting breast cancer[10]. The model's efficacy is tested by feeding it with new, labeled data, typically divided into training and testing sets through the train-test split method. About 75% of the data is used for constructing the model—known as the training set—while the remaining 25% serves as the test set to assess the model's performance. Post evaluation, the outcomes are analyzed to identify the algorithm that provides the highest accuracy and predictability for the presence of breast cancer. In the provided work, a comparative analysis of machine learning algorithms was conducted, focusing on key performance metrics like Accuracy, Precision, Recall, and F1 score. Figure 1 illustrates these comparisons in a comprehensive manner.

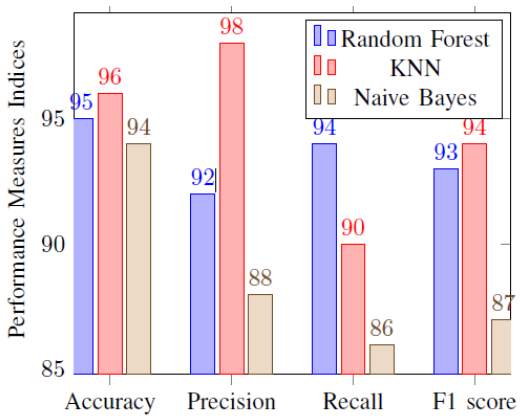


Figure 1: Comparative analysis of machine learning algorithms using performance metrics such as Accuracy, Precision, Recall, and F1 score.

Furthermore, a summary of different Machine Learning Techniques was compiled, as shown in Table 4. This table provides a concise overview of each algorithm, including their descriptions and a comparative analysis of their advantages and disadvantages.

Table 4: Summary of Machine Learning Techniques

Name of Algorithms	Descriptions	Advantages/Disadvantages
ANN	The output is generated through the combination of input and hidden layers.	Laborious operations and potentially subpar performance due to generic layered structure.
Decision Tree	Classification tree formed by nodes (variables) and leaves (decision outcomes).	Easy to interpret and fast learning process.
SVM	Identifies multiple hyperplanes in a high dimensional feature space and selects the best hyperplane for classifying input data into two classes.	Difficulty in handling large datasets.
Bayesian Networks	Makes estimates of probabilities rather than predictions [11].	Computationally expensive.

5 Results and Discussion

The performance of machine learning algorithms was evaluated using the Wisconsin Dataset. The models' performance was compared using metrics such as sensitivity, F1 score, confusion matrix, precision, and accuracy. Confusion matrices are particularly useful in assessing classification problems with two or more class types. They provide counts of True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP). Accuracy, the most common metric, is defined by the percentage of correctly predicted predictions from a specific sample size[4]. The accuracy rates for the Wisconsin dataset are displayed in Table 5 and Table 6

It was observed that while all classifiers demonstrated varying levels of accuracy, the Support Vector Machine (SVM) consistently outperformed others in the testing phase with an accuracy of 97.2%. Table V presents the confusion matrix, illustrating the performance of the classifiers in actual class conditions. According to the confusion matrix, SVM correctly predicted 556 out of 569 cases, including 201 actual cases of malignancy and 356 actual cases of benignity [11]. However, SVM also misclassified 11 benign cases as malignant and 1 malignant case as benign. This leads to SVM having higher accuracy compared to other classification methods [11]. The results indicate that SVM surpasses other classifiers with respect to sensitivity, precision, and F-Measure, all at 0.97%. In diagnosing malignant and benign classes in the Breast Cancer Wisconsin data, SVM is consistently superior to other classifiers [5].

Table 5: Confusion Matrix for Classifier Performance

Algorithm	Malignant	Benign
KNN	TP: 201, FP: 7	FN: 11, TN: 350
Logistic Regression	TP: 201, FP: 5	FN: 11, TN: 352
Random Forest	TP: 196, FP: 7	FN: 16, TN: 350
SVM	TP: 201, FP: 1	FN: 11, TN: 356

Table 6: Performance of Models During Testing Phase

Model	Recall	Precision	F1 Score	Accuracy
RF	94	92.2	93	93.74
K-nearest neighbour	91	98.3	94.3	96
Naive Bayes	86	89	86.4	94.5

6 Conclusion

Breast cancer remains the most prevalent cancer among women worldwide. Enhancements in diagnosis and prognosis are critical for health preservation. This study investigated two popular ML approaches for the categorization of Wisconsin Breast Cancer: Support Vector Machine (SVM) and Artificial Neural Networks (ANN). The findings indicate that SVM, with a precision of 97.5% and an Area Under the Curve (AUC) of 96.6%, outperforms all other algorithms, consistently delivering superior results in terms of diagnostic precision and accuracy for breast cancer. It is important to acknowledge the limitations of this study, which are primarily related to the confinement of the results to the WBCD database. Future work should consider applying the same methodologies to other databases to validate and compare the findings across different datasets. Furthermore, there is an intention to apply our and other ML algorithms to larger datasets with more disease classifications and additional parameters, aiming to achieve even more accurate results.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

Bhoomi Jain: Methodology, Software, Validation, Investigation, Data curation, Writing-original draft, visualisation.
Neetu Singla: Conceptualisation, Formal analysis, Resources, Writing-review & editing, supervision.

References

[1] B. Gayathri and C. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–5, IEEE, 2016.

[2] S. Kharya and S. Soni, "Weighted naive bayes classifier: a predictive model for breast cancer detection," *International Journal of Computer Applications*, vol. 133, no. 9, pp. 32–37, 2016.

[3] G. M. Hadi, "Benign and malignant breast cancer features based on region characteristics."

[4] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 114–118, IEEE, 2018.

[5] K. Sivakami and N. Saraswathi, "Mining big data: breast cancer prediction using dt-svm hybrid model," *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, vol. 1, no. 5, pp. 418–429, 2015.

[6] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[7] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 114–118, IEEE, 2018.

[8] M. F. bin Othman and T. M. S. Yau, "Comparison of different classification techniques using weka for breast cancer," in *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006: Biomed 2006, 11–14 December 2006 Kuala Lumpur, Malaysia*, pp. 520–523, Springer, 2007.

[9] B. Gayathri and C. Sumathi, "Mamdani fuzzy inference system for breast cancer risk detection," in *2015 IEEE international conference on computational intelligence and computing research (ICCIC)*, pp. 1–6, IEEE, 2015.

[10] Y. Buttan, A. Chaudhary, and K. Saxena, "An improved model for breast cancer classification using random forest with grid search method," in *Proceedings of Second International Conference on Smart Energy and Communication: ICSEC 2020*, pp. 407–415, Springer, 2021.

[11] S. Goyal, M. Sinha, S. Nath, S. Mitra, and C. Arora, "Breast cancer detection using machine learning," in *Communication, Software and Networks: Proceedings of INDIA 2022*, pp. 613–620, Springer, 2022.

JCMM Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 2 Issue 6

Article Number: 230110

Enhanced Shopping Experiences: The Role of RFID Technology in Smart Carts

Ipshita Datta*[1], Anjali Garg[1], Anu Tonk[1], and Pankaj Rakheja[2]

[1]Department of Multidisciplinary Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017
[2]Department of Computer Science and Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

Abstract

Over the past decade, the Internet has revolutionized the way individuals shop, with the rise of e-Commerce allowing consumers to buy products online through platforms like Amazon and Flipkart. However, traditional physical markets are still growing and offer a personal shopping experience. This project aims to create a Smart Shopping Cart that can be used in supermarkets to address common customer issues such as waiting in long queues for billing and overspending. This technology reduces the burden on cashiers by automatically calculating the total cost of the items in the cart. The Smart Cart provides a fast and efficient service, reducing congestion at the checkout counter and saving time and effort during the accounting process. The items in the cart are scanned for price and barcode details and the cost of each item is calculated and sent to an Arduino device. In traditional Indian markets, customers place items in a cart and then stand in a queue for billing, but this Smart Cart eliminates the need for this step.

Keywords: RFID Tag; LCD, RFID Reader; Barcode Scanner; WiFi; Central Charging Unit

1 Introduction

Humans have consistently invented new technologies to satisfy their needs and requirements since the dawn of mankind. The primary motive behind the development of new inventions is to gain more independence, resulting in the improvement of tasks and making routine ones simpler and faster [1]. One everyday activity that requires significant effort is shopping. A store is a place where individuals acquire their daily necessities, ranging from basic items like food and garments to electronic gadgets [2]. Traditionally, customers choose their desired items, place them in their shopping baskets or trolleys, and then face prolonged queues at the checkout counters for billing. The payment process at these counters is often frustrating and time-consuming, leading to significant congestion and wait times [3]. According to research, people spend approximately 1.4 hours shopping in stores, and many will leave if the lines are too long [4]. The proposed smart basket system aims to enhance the efficiency of in-store shopping. This innovation involves a handbasket equipped with RFID labels to track the items in the cart, along with an RFID reader and an LCD screen. This setup enables customers to view the cost of each item and the total cost of all items in the cart, thereby saving time for customers and reducing the staffing needs in shopping complexes, making the shopping experience more convenient and efficient. RFID (Radio-Frequency Identification) is an emerging technology that has recently garnered significant attention in research communities due to its advantages over existing systems. RFID technology facilitates the exchange of information between labels and readers without the need for a direct line of sight, covering distances of several meters, depending on the type of tag used.

*Corresponding author: ipshitadatta1@gmail.com
Received: 14 November 2023; **Revised:** 30 November 2023; **Accepted**: 04 December 2023; **Published:** 31 December 2023

DOI: 10.57159/gadl.jcmm.2.6.230110.

In RFID systems, data communication is conducted via radio waves, and the labels are consistently scanned or collected. The technology involves storing information on a device akin to a smart card, offering numerous advantages over other commonly used identification systems. The key strengths of RFID technology include its ability to rapidly and accurately read data from tags without requiring physical contact or line-of-sight positioning, which significantly enhances efficiency in various applications such as inventory management, asset tracking, and smart shopping systems [5–8]. The integration of RFID technology in shopping carts can streamline the shopping experience by automatically identifying and tallying items as customers add them to their carts, mitigating the need for manual scanning at checkout counters. This advancement not only expedites the billing process but also contributes to reducing queue times and improving overall customer satisfaction in retail environments. Figure 1 schematically represents the working of RFID technology.

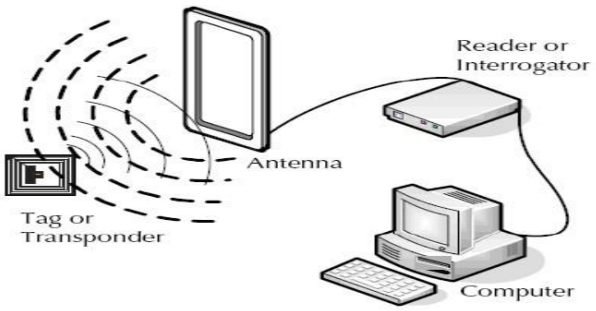


Figure 1: Working of RFID system

The primary objective of this project is to develop an enhanced shopping experience by integrating RFID technology into a shopping cart. The focus is on embedding RFID-based strategies within the cart's functionality, employing RFID cards to secure product acquisition within shopping complexes. When an item is placed in the cart, its cost is promptly displayed on the LCD, contributing to the cumulative total. Conversely, the removal of an item results in an immediate deduction from the total amount. The central aim is to establish a shopping environment that is more user-friendly, streamlining and expediting the shopping process compared to the current paradigm. This innovation is expected to significantly reduce waiting times at billing counters, offering considerable time-saving benefits to customers.

2 Related Work

2.1 Analysis of RFID from printed reviews

An examination delving into the benefits of tagging items with RFID labels for routine operations and business value in high-quality retail settings reveals significant advantages. The findings indicated that the overall stock accuracy improved by an additional 27% when using RFID technology [9]. Several studies [10, 11, 6] have also compared the efficiency of managing inventory with RFID tags versus traditional barcode tag readers. It was observed that with RFID, the process of assessing an equivalent number of articles required only two hours, whereas using a conventional barcode scanner necessitated fifty-three hours. This dramatic reduction in time highlights the superiority of RFID in terms of operational efficiency. The ability of RFID to read multiple tags simultaneously without direct line-of-sight [12] contributes to this efficiency, making it an ideal solution for fast-paced retail environments. Implementing RFID not only enhances inventory management but also translates into better customer service by reducing the likelihood of stockouts and inaccuracies in inventory records [13, 14, 9]. Consequently, RFID technology represents a significant advancement over traditional barcode systems, offering tangible benefits in terms of time savings and improved accuracy in retail settings.

2.1.1 RFID Usage Challenges

While Radio-Frequency Identification (RFID) technology offers notable advantages in terms of speed and efficiency, it cannot entirely replace traditional barcode technology due to its comparatively higher cost and lower precision [15]. Nevertheless, prominent companies in various sectors such as CVS, Tesco, Prada, Benetton, Walmart, and Procter & Gamble are actively adopting and investigating the impact of RFID technology on their operations. The potential for RFID adoption extends beyond these industries, but it is essential to comprehensively understand its development and various facets to mitigate potential challenges. The effective implementation of an RFID system requires considerable energy and commitment, as this technology faces various implementation challenges. The primary challenges include issues related to scalability, global integration, regulatory compliance, and cost considerations.

For instance, each shopping cart needs to be outfitted with a device that contains an RFID sensor. Furthermore, every product within the shopping center must be tagged with RFID labels [3]. As customers place items into their carts, the RFID sensor identifies the corresponding data, and the cost of each product is recorded in the storage unit of the cart. As additional items are added, their prices are cumulatively calculated to determine the total amount. This

entire computation process takes place within the cart itself. Details about the items, including their prices, are displayed on an LCD screen attached to the cart. As customers near the payment center, comprehensive invoice information is transmitted to a personal computer through encrypted radiofrequency modules. This step ensures that the billing process is seamless and efficient. However, the widespread implementation of this system poses significant logistical and financial challenges. Ensuring that all products are accurately tagged and that each cart is properly equipped and maintained requires substantial investment and coordination [16]. Additionally, the system must be designed to handle the vast quantity of data generated and to integrate smoothly with existing retail management systems. Despite these challenges, the potential benefits of RFID in enhancing the shopping experience and streamlining retail operations make it a promising area for further development and innovation.

2.2 Comparative Analysis of Existing Model

The existing smart cart models employ various technologies and offer a range of features, advantages, and disadvantages as summarized in Table 1.

Table 1: Comparison analysis of existing models

Smart Cart Model	**RFID Technology Used**	**Features**	**Advantage**	**Disadvantages**
Amazon Dash [17–19] Cart	Ultra-wideband (UWB) and RFID	Automatically detects products and adds them to the cart, checkout-free experience, built-in scale for measuring weight.	Removes the need for cashiers and conventional checkout lines, provides real-time information about the total cost of items purchased, and saves time.	Limited to Amazon Fresh grocery stores only, UWB technology has a limited range.
Kroger Smart Cart [20**?** –23]	RFID	Scans and weighs items automatically, allows shoppers to pay directly from the cart, and has a built-in touchscreen for adding or removing items.	Saves time, removes the need for cashiers and conventional checkout lines, and allows shoppers to keep track of the total cost of their items.	Currently available in only one location in the U.S.
Wheelys Moby Mart [24–27]	RFID	A self-driving convenience store that uses RFID to track items taken by shoppers, and automatically charges their account when they leave the store.	Convenient and accessible, removes the need for cashiers and conventional checkout lines and provides real-time inventory management.	Limited to certain locations and may not work for all types of products.
Intelligent Cart [28–30]	RFID	Automatically scans and weighs items, tracks a shopper's location in the store, and provides recommendations based on their shopping history.	Saves time, provides personalized recommendations, and eliminates the need for cashiers.	Limited availability and may not work for all types of products.
Scandit Smart Grocery Cart [31, 32]	RFID	Scans and weighs items, allows shoppers to pay directly from the cart, and provides real-time information about the total cost of items purchased.	Saves time, removes the need for cashiers and conventional checkout lines, and provides real-time inventory management.	Limited availability and may not work for all types of products.

Table 1 presents the comparison of existing hardware with a few features matching the proposed model. It is important to note that such hardware does not exist at the national level, and at the international level, the cost of the product is too high with fewer features as compared to the proposed model. The proposed model offers a lot of additional features

with very less cost which will be affordable for many retailers. A reseach works have demnstrated smart shopping cart leveraging RFID technology for product identification and an automated billing system for calculating total item costs within the cart [33, 34]. The framework generally employs an Arduino microcontroller to handle data from the RFID reader and present the cumulative cost on an LCD display. Empirical tests utilizing RFID reader and tags demonstrated high accuracy and efficiency. Yet another research introduced a smart shopping cart harnessed with RFID technology for product recognition and an automatic billing system for computing total cart costs. The setup integrates an RFID reader, microcontroller, LCD display, and even includes a barcode scanner to enhance product identification. Thorough evaluation involving RFID tags and barcode scanning affirmed its proficiency in product identification and cost computation [35]. Another research work focused on developing an automated billing system. This system leverages RFID (Radio-Frequency Identification) technology and cloud computing to streamline the billing process. Their work aims to enhance the efficiency and accuracy of billing systems, particularly in retail environments, by automating the identification and calculation of product prices. The integration of RFID with cloud computing in their proposed system suggests a modern approach to handling transactions, aiming to reduce manual effort and errors associated with traditional billing methods [36]. In these studies, researchers have innovatively harnessed RFID technology in smart shopping cart systems, with each approach contributing novel insights and benefits to enhance the shopping experience.

2.3 Research Gaps

Despite the technological advancements in RFID technology, there is a critical need to revitalize the structural framework of this technology to ensure scalability in line with evolving requirements. Current literature has significantly focused on RFID anti-collision protocols to enhance system performance. However, empirical studies evaluating the benefits of the Cipher Isolation approach remain limited [37]. A substantial body of research has been dedicated to assessing the performance of RFID systems, but there are still challenging issues that need addressing to improve the overall functionality of these systems. One notable gap in the literature is the limited exploration of fault detection and uncertainty handling in RFID data. Addressing real-world uncertainties associated with RFID data is essential for achieving more semantically robust objectives for broader applications. Moreover, as indicated in the initial part of this literature review, the volume and speed of data generation in RFID systems are set to outpace the capabilities of the current technological infrastructure. It is therefore crucial not only to propose designs that can manage advanced RFID data but also to address existing challenges in RFID systems, taking into account real-world vulnerabilities. This pressing need highlights the importance of pushing this area of concern to the forefront and striving for comprehensive advancements in RFID technology.

3 Methods

3.1 Proposed Methodology

The proposed framework detailed in this thesis is divided into two main sections, as illustrated in Figures 2 and 3. These figures present the block diagrams of the envisioned system, showcasing how the RFID reader interfaces with an Arduino Nano, which in turn connects to both an LCD and a WiFi module, thus facilitating the transmission of billing details to the central billing unit. Within the transmitter section, as depicted in Figure 2, the reader detects RFID tags on the products and conveys the acquired data to the Arduino Nano. The Arduino Nano then compares this data with its internal database and displays the result on the connected LCD screen. This part of the system is responsible for providing immediate feedback to the consumer regarding the cost of individual items as well as the running total of their shopping cart. The WiFi transmitter component is tasked with sending the data showcased on the LCD screen to the central billing unit. This information is captured by a WiFi receiver at the central billing unit. The central billing unit, as shown in Figure 3, then processes the received data to generate a printed record of the transaction [38]. This methodology underscores the seamless integration of different components to create a holistic system that enhances the shopping experience by automating the billing process and minimizing the time consumers spend at the checkout counter.

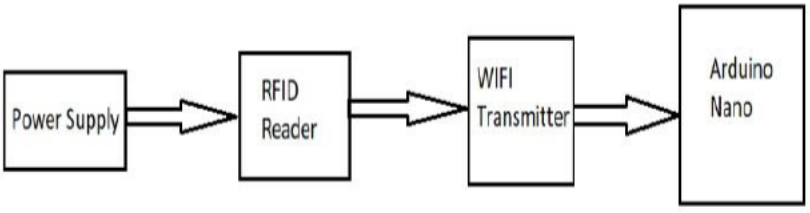


Figure 2: Block diagram of the transmitter section of the proposed system.

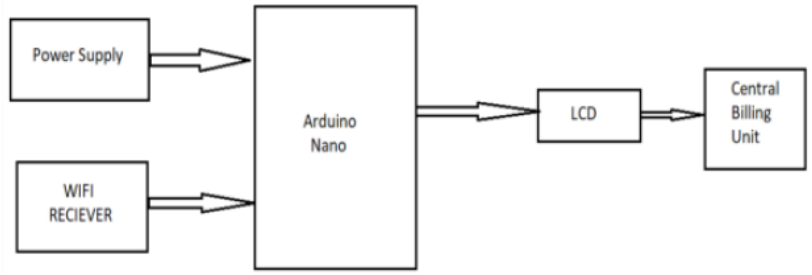


Figure 3: Block diagram of the receiver section of the proposed system.

4 System Design

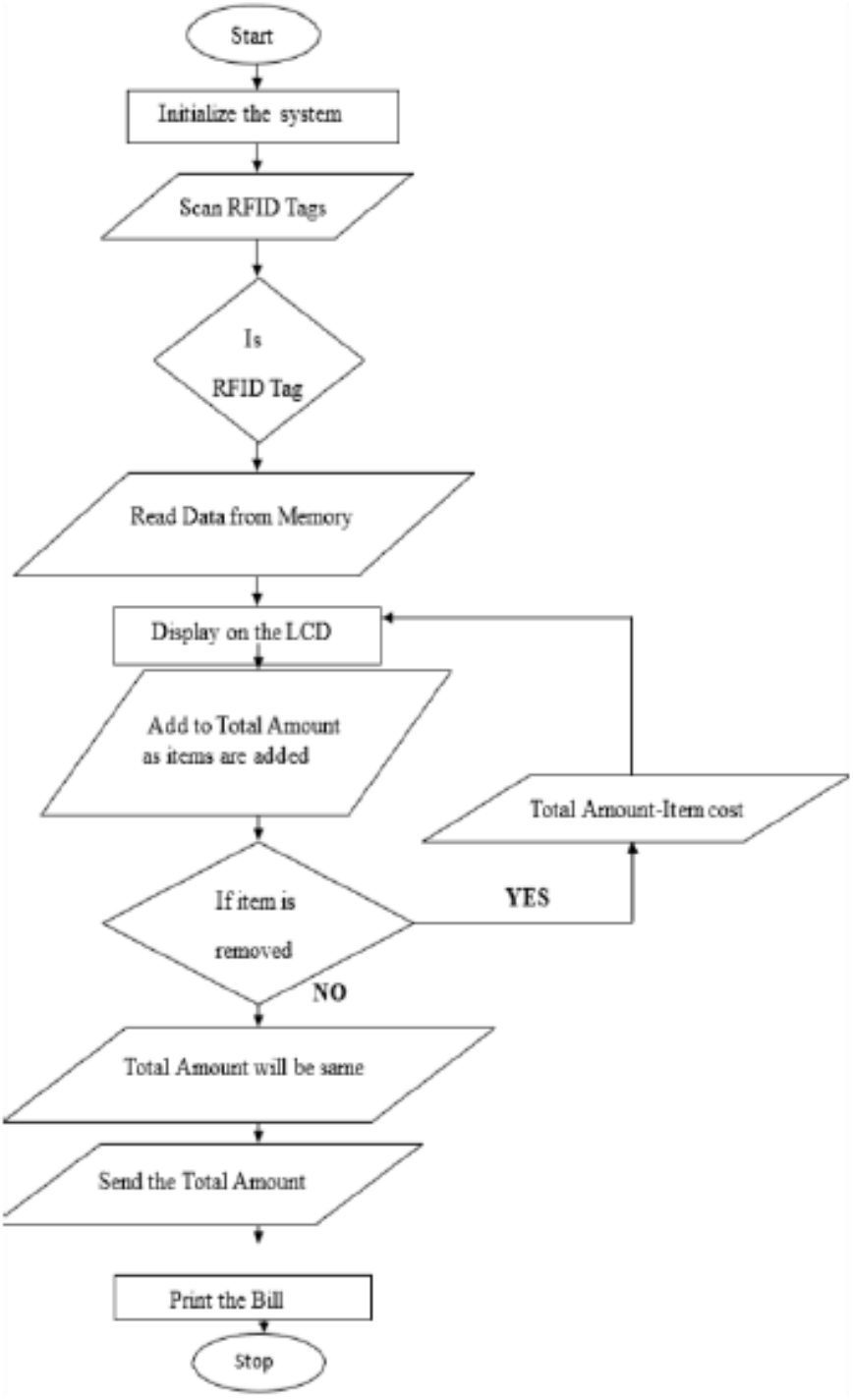


Figure 4: Flowchart illustrating the operation of the proposed system.

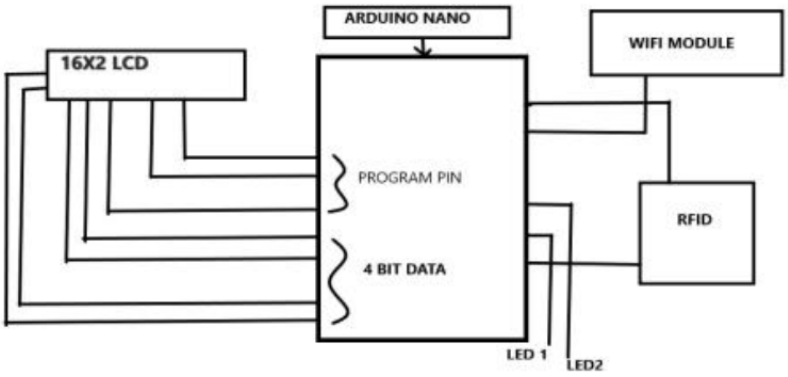


Figure 5: Schematic design of the proposed system.

Figure 4 presents the flowchart that illustrates the working mechanism of the proposed system. Figure 5 details the schematic design of the system. The system's hardware setup is as follows: The Arduino Nano is powered by a 5V supply via USB from a computer. The LCD's input power pins are connected in the following manner: Pin 1 (GND) to ground, Pin 2 (VCC) to a 5V power source, and Pin 3 (VEE) to a 1k ohm resistor. The data command pins, Pin 4 (RS) and Pin 6 (EN), are connected to Arduino's digital pins D-2 and D-3, respectively. The data pins, D4 to D7, are connected to Arduino's digital pins D-4 to D-7 respectively. For the RFID module, it is supplied with 3.3V power through a 5V voltage regulator, with its communication pins, RST, SDA, MOSI, MISO, and SCK, connected to Arduino's digital pins D-9 to D-13 respectively. The Wi-Fi module's RX and TX pins are interfaced with the corresponding RX and TX pins on the Arduino, ensuring data transmission to the central billing system. This configuration is intended to ensure seamless connectivity and functionality among the various components, leading to efficient operation of the overall system.

4.1 System Components

4.1.1 Arduino NANO

The Arduino Nano is a small, complete, and breadboard-friendly microcontroller board based on the ATmega328 [39]. It is devoid of a DC power jack and is instead powered via a Mini B USB cable. The Nano, developed by Arduino, is designed for convenience and ease of integration into various projects, including those that are space-constrained.

4.1.2 RFID Module

The RFID module employed in this project is based on the EM18 RFID Module. It provides a convenient and time-saving means of reading data through microcontrollers [40]. The module requires a simple serial connection for interface and is compatible with other boards through +5V and GND header wires. It includes an onboard 5V voltage regulator, allowing for a 9-15V DC adapter power supply. The board is equipped with a red LED to indicate power status and a green LED with buzzer for card or tag detection. A selection jumper is available for switching between output formats, including a Weigand 26 O/P. It is important to note that selecting a different mode will change the protocol, affecting the ability to read data through standard serial protocols. The following are the features:

- The RFID module is an economical option for reading passive RFID transponder tags. It operates at a frequency of 125 kHz with a 9600 bps serial interface and is capable of reading EM 4100 compatible transponders [40].
- The module offers a read distance of up to 6-10 cm for cards and 5 cm for key tags, featuring onboard LED and buzzer indications.
- It provides output through both serial UART from pin headers (TTL CMOS) and RS-232 (DB9), facilitating easy interfacing with a range of devices.

4.2 ESP 8266

The ESP8266 is a low-cost Wi-Fi microchip with full TCP/IP stack and microcontroller capability, developed by Espressif Systems [41]. Since its release in 2014, it has gained immense popularity in the IoT community for its efficiency, small size, and integrated Wi-Fi functionalities. The ESP8266 is an excellent choice for IoT projects due to its versatility, power, and cost-effectiveness.

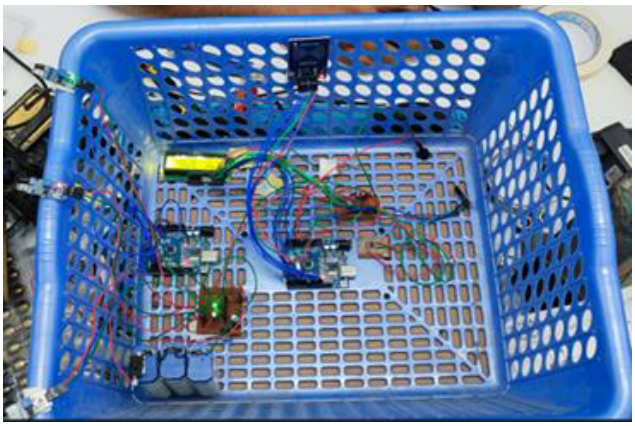
4.3 Software Used

Remote XY is a versatile platform designed for creating custom mobile applications and web interfaces for the remote control of microcontrollers and embedded systems using Wi-Fi, Bluetooth, and other communication protocols [42, 43].

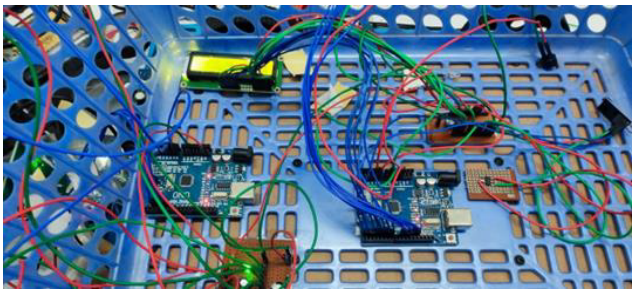
It enables users to control and monitor their projects remotely from smartphones, tablets, or computers. The platform offers a user-friendly drag-and-drop interface for designing custom user interfaces. It also includes a variety of pre-built widgets for controlling hardware components such as buttons, switches, sliders, and gauges. One of the distinctive features of Remote XY is its compatibility with a wide range of microcontrollers, including Arduino, Raspberry Pi, ESP32, ESP8266, among others. Users have the liberty to select the communication protocol that best fits their project requirements, such as Wi-Fi, Bluetooth, or USB. This adaptability has made Remote XY a preferred choice for both hobbyists and professionals who aim to remotely control and monitor their projects [42].

5 Results and Discussion

The prototype of the smart shopping basket, as depicted in Figures 6a and 6b, demonstrates the successful integration of RFID technology with the essential hardware components. The RFID system is adept at detecting and measuring the weight of items placed within the basket. The integrated LCD screen provides a real-time display of each item's cost, facilitating the shopper's ability to make informed decisions based on their running total.



(a) Top view of the shopping basket prototype



(b) Close-up view of the RFID and electronic components

Figure 6: Hardware connections of smart shopping basket prototype

This innovation addresses common consumer pain points, such as the time spent at checkout queues and the difficulty of tracking spending during shopping. Compared to traditional shopping methods, the smart basket offers an enhanced user interface that informs the shopper of their total expenditure in real time. This feature not only improves the shopping experience but also aids in budget management. However, the technology's reliance on the visibility and orientation of RFID tags could pose challenges in ensuring consistent performance. The current prototype's capability to measure item weight and display pricing information instantly sets it apart from conventional shopping methods. Future iterations of the smart basket could include more advanced features, such as integration with a store's inventory management system, personalized shopping recommendations based on consumer purchase history, and even the ability to sync with mobile payment systems for a seamless checkout process. The effectiveness of RFID in reducing checkout time has been validated through empirical testing, showing a high degree of accuracy and efficiency. Nonetheless, considerations must be made regarding the scalability of the system and the potential need for retailers to update their inventory practices to fully leverage the capabilities of RFID technology. In conclusion, the smart shopping basket prototype represents a promising direction in retail innovation. Continued development and integration with broader retail systems could redefine the shopping experience, offering benefits to both consumers and retailers alike.

6 Conclusion

Progress in science and technology is a continuous and dynamic process, with novel devices and technologies being constantly conceptualized and developed. This project has contributed to such advancements by leveraging RFID technology to enhance the shopping experience in retail centers. The employment of RFID technology served not only to ensure secure product access but also to bolster surveillance efficiency, addressing significant concerns in retail operations. The primary objective of this endeavor was to establish an automated billing system or central charging mechanism, aimed at eliminating the need for customers to queue at payment counters after their purchases. This innovation has resulted in a swifter billing process, streamlining and simplifying the overall transactional procedure.

Furthermore, the system's ability to detect theft attempts by fraudulent customers enhances its reliability and appeal to both customers and merchants, taking the shopping experience to a new level of efficiency and security. Looking ahead, it is recommended to explore the use of more powerful RFID readers with increased capacity to handle a greater number of products in the cart. Such advancements will likely further enhance the system's efficiency and reliability, thereby continuing the trend of innovation in retail technology and customer service.

Declaration of Competing Interests

The authors declare the following competing interests in relation to the development and discussion of smart cart technology utilizing Radio-Frequency Identification (RFID): Research Funding: The research and development of the smart cart using RFID technology discussed in this context have received partial funding from Thr Northcap university, which could affect the interpretation of the outcomes. Personal Interests: The authors acknowledge their personal belief in the transformative potential of RFID technology in enhancing smart cart functionality, which might influence my interpretations and opinions. Authors affirm that these competing interests do not compromise the objectivity or integrity of the information presented herein. Thus, the authors believe to have maintained transparency and offer this declaration to ensure transparency and allow readers to assess her perspectives with full awareness of potential biases.

Funding Declaration

The research and development activities of the implementation of RFID technology in the smart cart project have received financial support from the Vice-Chancellor Innovation Fund, NortCap University. This funding has been instrumental in conducting experiments, acquiring the necessary equipment, and supporting the personnel involved in this project.

Author Contribution

Ipshita Datta: Conceptualization, Methodology, Data Curation, Software; **Anjali Garg**: Supervision, Writing - Review and Editing; **Anu Tonk**: Supervision, Writing - Review and Editing; **Pankaj Rakheja**; Supervision, Writing - Review and Editing

References

[1] Z. Tekic and D. Koroteev, "From disruptively digital to proudly analog: A holistic typology of digital transformation strategies," *Business Horizons*, vol. 62, no. 6, pp. 683–693, 2019.

[2] A. Rossolov, H. Rossolova, and J. Holguín-Veras, "Online and in-store purchase behavior: shopping channel choice in a developing economy," *Transportation*, vol. 48, no. 6, pp. 3143–3179, 2021.

[3] S. Suraj, V. Guruprasad, U. R. Pranava, and P. S. Nag, "Rfid based wireless intelligent cart using arm7," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 5, no. 8, 2016.

[4] F. F. Dias, P. S. Lavieri, S. Sharda, S. Khoeini, C. R. Bhat, R. M. Pendyala, A. R. Pinjari, G. Ramadurai, and K. K. Srinivasan, "A comparison of online and in-person activity engagement: The case of shopping and eating meals," *Transportation Research Part C: Emerging Technologies*, vol. 114, pp. 643–656, 2020.

[5] R. Want, "An introduction to rfid technology," *IEEE pervasive computing*, vol. 5, no. 1, pp. 25–33, 2006.

[6] X. Zhu, S. K. Mukhopadhyay, and H. Kurata, "A review of rfid technology and its managerial applications in different industries," *Journal of Engineering and Technology Management*, vol. 29, no. 1, pp. 152–167, 2012.

[7] B. Nath, F. Reynolds, and R. Want, "Rfid technology and applications," *IEEE Pervasive computing*, vol. 5, no. 1, pp. 22–24, 2006.

[8] M. Kaur, M. Sandhu, N. Mohan, and P. S. Sandhu, "Rfid technology principles, advantages, limitations & its applications," *International Journal of Computer and Electrical Engineering*, vol. 3, no. 1, p. 151, 2011.

[9] T.-J. Fan, X.-Y. Chang, C.-H. Gu, J.-J. Yi, and S. Deng, "Benefits of rfid technology for reducing inventory shrinkage," *International Journal of Production Economics*, vol. 147, pp. 659–665, 2014.

[10] K. Michael and L. McCathie, "The pros and cons of rfid in supply chain management," in *International Conference on Mobile Business (ICMB'05)*, pp. 623–629, Ieee, 2005.

[11] G. White, G. Gardiner, G. P. Prabhakar, and A. Abd Razak, "A comparison of barcoding and rfid technologies in practice," *Journal of information, information technology and organizations*, vol. 2, 2007.

[12] Y. Ma, N. Selby, and F. Adib, "Minding the billions: Ultra-wideband localization for deployed rfid tags," in *Proceedings of the 23rd annual international conference on mobile computing and networking*, pp. 248–260, 2017.

[13] S. Goyal, B. C. Hardgrave, J. A. Aloysius, and N. DeHoratius, "The effectiveness of rfid in backroom and sales floor inventory management," *The International Journal of Logistics Management*, vol. 27, no. 3, pp. 795–815, 2016.

[14] B. C. Hardgrave, J. A. Aloysius, and S. Goyal, "Rfid-enabled visibility and retail inventory record inaccuracy: Experiments in the field," *Production and Operations Management*, vol. 22, no. 4, pp. 843–856, 2013.

[15] K. G. Devi, T. A. Kaarthik, N. K. Selvi, K. Nandhini, and S. Priya, "Smart shopping trolley using rfid based on iot," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 3, 2017.

[16] M. Attaran, "Rfid: an enabler of supply chain operations," *Supply Chain Management: An International Journal*, vol. 12, no. 4, pp. 249–257, 2007.

[17] M. F. Farah and Z. B. Ramadan, "Disruptions versus more disruptions: How the amazon dash button is altering consumer buying patterns," *Journal of Retailing and Consumer Services*, vol. 39, pp. 54–61, 2017.

[18] M. F. Farah and Z. B. Ramadan, "Viability of amazon's driven innovations targeting shoppers' impulsiveness," *Journal of Retailing and Consumer Services*, vol. 53, p. 101973, 2020.

[19] C. Livingstone and I. Knezevic, "From online cart to plate: What amazon's retail domination means for the future of food," *Journal of Agriculture, Food Systems, and Community Development*, vol. 9, no. 4, pp. 311–329, 2020.

[20] R. Aydin[1], M. B. Jaweesh, A. AlHunaiyyan, and M. Mansour, "Check for updates assessing the implementation of a smart cart in a supermarket using a simulation model," in *Towards a Smart, Resilient and Sustainable Industry: Proceedings of the 2nd International Symposium on Industrial Engineering and Automation ISIEA 2023*, vol. 745, p. 84, Springer Nature, 2023.

[21] M. A. Sarwar, Y.-A. Daraghmi, K.-W. Liu, H.-C. Chi, T.-U. İk, and Y.-L. Li, "Smart shopping carts based on mobile computing and deep learning cloud services," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2020.

[22] H. Abbu and P. Gopalakrishna, "Digital transformation powered by big data analytics: the case of retail grocery business," 2022.

[23] R. Aydin, M. B. Jaweesh, A. AlHunaiyyan, and M. Mansour, "Assessing the implementation of a smart cart in a supermarket using a simulation model," in *International Symposium on Industrial Engineering and Automation*, pp. 84–95, Springer, 2023.

[24] M. Cometti, D. Fredette, A. E. Panek, and M. Radley, "From markets to tech: Governmental initiatives, solutions, and responses to food insecurity," *Journal of Public Affairs*, vol. 19, no. 3, p. e1860, 2019.

[25] C. Lyster, "Disciplinary hybrids: Retail landscapes of the post-human city," *Architectural Design*, vol. 89, no. 1, pp. 100–105, 2019.

[26] K. Tetiana, "Kopchykova inna," *Strategies for sustainable socio-economic development and mechanisms their implementation in the global dimension*, p. 186, 2019.

[27] E. Roderick, "Mobile markets/grocery stores," *Increasing Access to Food: A Comprehensive Report on Food Supply Options*, p. 114, 2019.

[28] A. Pimple, "Intelligent cart," *Methodology*, vol. 6, no. 04, 2019.

[29] Y. Kambayashi, Y. Harada, O. Sato, and M. Takimoto, "Design of an intelligent cart system for common airports," in *2009 IEEE 13th International Symposium on Consumer Electronics*, pp. 523–526, IEEE, 2009.

[30] T. Lekhaa, S. Rajeshwari, J. A. Sequeira, and S. Akshayaa, "Intelligent shopping cart using bolt esp8266 based on internet of things," in *2019 5th International Conference on Advanced Computing & Communication Systems*, pp. 758–761, IEEE, 2019.

[31] M. Ritchie, T. Longino, D. Garza, and A. Katranci, "Optimized automated checkout process for major food retailers," 2021.

[32] S. Parameswaran and N. J. Zaninovich, "Future of self checkout,"

[33] P. Chandrasekar and T. Sangeetha, "Smart shopping cart with automatic billing system through rfid and zigbee," in *International Conference on Information Communication and Embedded Systems (ICICES2014)*, pp. 1–4, IEEE, 2014.

[34] P. Aryan, P. Pise, S. Tamhane, R. Patil, and K. Pittulwar, "Smart shopping cart with automatic billing system through rfid and bluetooth," *International Journal of Emerging Technology and Computer Science*, vol. 1, no. 02, pp. 73–76, 2016.

[35] A. K. S. Al-Hakmani and A. Sajan, "Design and implementation of a smart shopping trolley using rfid technology," *Journal of Student Research*, 2019.

[36] D. Sinha, K. Cottur, K. Bhat, and C. Guruprasad, "Automated billing system using rfid and cloud," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, vol. 1, pp. 1–6, IEEE, 2019.

[37] S. Sharma and T. Kumar, "Smart shopping cart," *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 8, pp. 437–442, 2020.

[38] K. Ambekar, V. Dhole, S. Sharma, and T. Wadekar, "Smart shopping trolley using rfid," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 4, no. 10, pp. 3875–3877, 2015.

[39] J. Suryaprasad, P. B. O. Kumar, D. Roopa, and A. K. Arjun, "A novel low-cost intelligent shopping cart," in *2014 IEEE*, 2014.

[40] A. Karmouche and Y. Salih-Alj, "Aisle-level scanning for pervasive rfid-based shopping applications," in *2012 International Conference on Computer Systems and Industrial Informatics*, pp. 1–4, IEEE, 2012.

[41] P. Shah, J. Jha, N. Khetra, and M. Zala, "A literature review on improving error accuracy and range based on rfid for smart shopping," *International Journal for Scientific Research & Development*, vol. 3, no. 10, pp. 2321–0613, 2015.

[42] H. A. bAmeer, M. A. Mohammedali, L. A. Abdul-Rahaim, and I. H. Al-Kharsan, "Design of surgical arm robot based on cloud computing," in *2022 5th International Conference on Engineering Technology and its Applications (IICETA)*, pp. 289–293, IEEE, 2022.

[43] M. Karabatak, T. Mustafa, and C. Hamaali, "Remote monitoring real time air pollution-iot (cloud based)," in *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–6, IEEE, 2020.

Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AAN PUBLISHING

Volume 2 Issue 6

Article Number: 230107

Bond Strength of Substrate With Repair Material for Masonry Structures: Scientometric Analysis and Literature Review

Khyati Saggu*, Shilpa Pal, and Nirendra Dev

Department of Civil Engineering, Delhi Technological University, New Delhi, India 110042

Abstract

Masonry infrastructure repair is a daunting challenge. The short intervention and limited resources have led to the loss of novel rendering and finishing materials. In the present study, the bibliometric analysis is conducted for period 2003 to 2022. Co-cited reference analysis, keywords, cluster, and temporal evolution were used for analysis using Citespace software. Additionally, summary of suitable content showed that major challenges in this field are disparities in the properties of old mortar and new materials and the lack of documents to understand nature and methodology of construction. Based on the analysis and content review, a suggestive technique is proposed inculcating four aspects: Finite element analysis, preparing numerical models, improvised techniques for bonding mechanisms and understanding mechanical attributes.

Keywords: Repair Material; Mortar; Bond Strength; Co-Citation Analysis; Characterization

1 Introduction

Historic structures are the best expression of the skill, ethos, and antiquity of their builders. For centuries, the most long-lived and impressive monuments have been evidence of true heritage spirit. With the advent of technology and deteriorating environmental conditions, it has been observed that the old infrastructure around the world has reached the stage of sheer disorder. This calls for substantial efforts from all stakeholders to render our waning infrastructure back to an operative and benign state. Therefore, the importance of the structural conservation of old constructions is to value history over progress as they signify the past of a specific region and apprise people about the construction and structural convention [1]. Keeping in view this much-needed aspect, many countries have come together to collaborate to fund the 1977 World Heritage Fund to safeguard world culture and Natural heritage that holds outstanding global worth. World Heritage Fund amounts to 5.9 billion dollars for the biennium 2022-23. The huge amount while the world face crisis due to epidemic shows the awareness and the interest of the nations to preserve their history. The culminating collaboration in USD by nations over the last 11 years is presented in Figure 1 [2]. Historic preservation is a vast domain that categorizes the major tasks as a visual inspection that reflects the building's prevailing condition, viability analysis i.e., analysis of the building based on life-safety issues, compliance with codes and anticipated longevity, context assessment, by comparison of the building with the similar building of that era, and design of rehabilitated building structure, and repair materials for historic structures [3, 4]. Though the planning of the intervention for many historical structures is significant the execution area still lacks. This is due to improper evaluation of the site; lack of resilience in our restoration materials, incapability of repairs to regulate more deterioration and an absence of lucrative placement technologies [3, 5]. This is well-curtained that the efficiency of mortar repairs depends on the quality and behaviour of the interface between the repair material and the substrate material in existing structures. Therefore, the researchers are significantly working on the reparation materials to understand their behaviours to carry out restoration and repair [4, 5].

*Corresponding author: khyatisaggu_2k21phdce501@dtu.ac.in
Received: 14 November 2023; **Revised:** 26 November 2023; **Accepted**: 04 December 2023; **Published:** 31 December 2023

DOI: 10.57159/gadl.jcmm.2.6.230107.

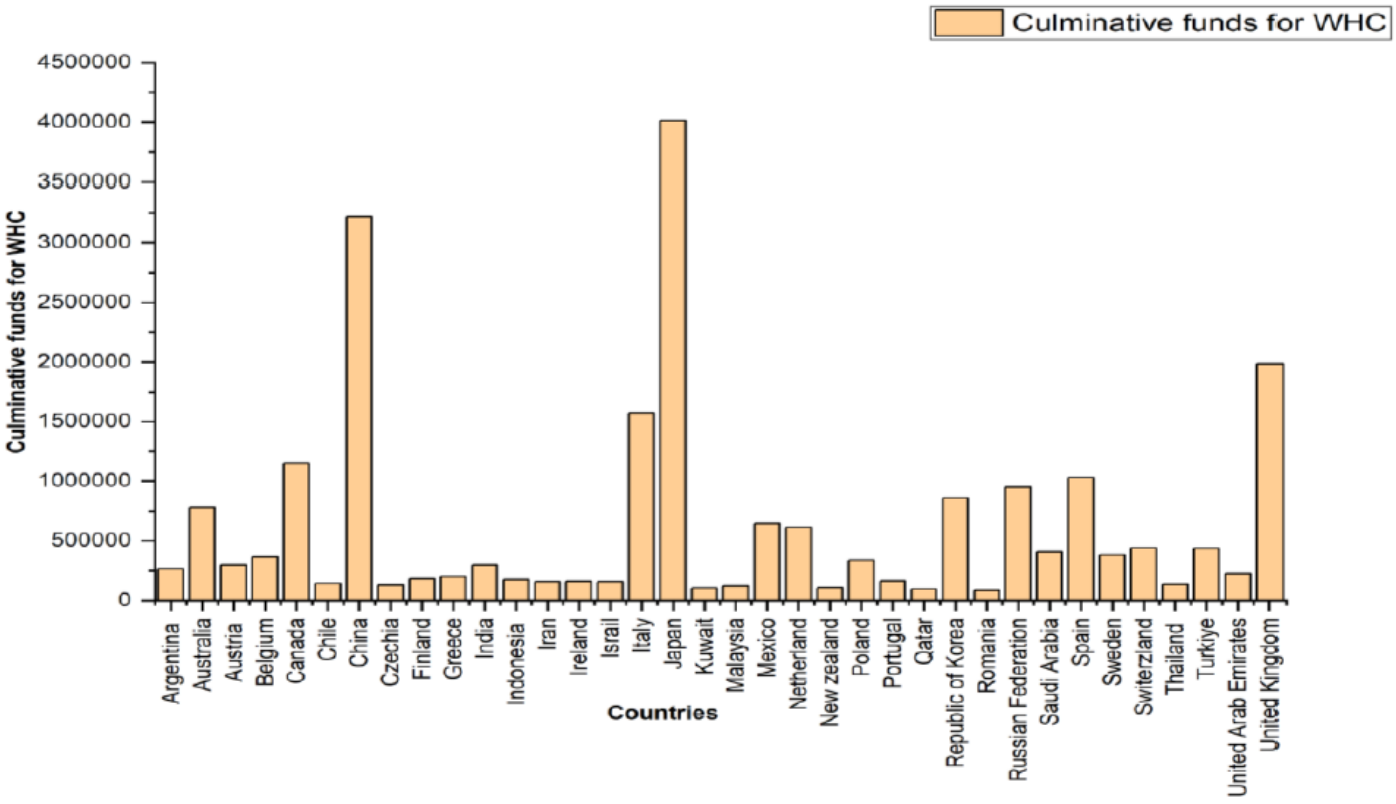


Figure 1: Culminative WHC Funds of different nations in US Dollars

Though the literature shows that the basic techniques and materials of masonry have remained untouched through the ages still there exists a dilemma in employing proper methods to repair them [6]. This concern has come to the limelight recently but still exists a few gaps that need to be addressed. There is a need to acquire a thorough comprehension of recent studies about the bond strength of substrate and repair material for ancient structures and to recognize research disparities. Therefore, the subsequent study is executed to conduct Systematic Literature Review (SLR). The SLR is an essential requirement in the present scenario as it enhances the information about the subject to be researched, and acts as the source of guidance for the selection of a suitable domain. This also helps researchers to identify problems initially and make amendments in future studies. It also provides insight into the state of knowledge about prospects to synthesize information from novel research, evaluate the degree of consensus or its dearth in the field, elaborate forthcoming opportunities, and tasks. About the information provided above, the subsequent questions are recognized in the paper.

Q1. What is the consideration for the bond between the substrate of the existing material and the repair material? What is the basis of the selection of suitable repair material?

Q2. What is the basis of research scholars' distribution working in this domain along with details of their institutes and nationalities?

Q3. What are the typical field classification and related priorities?

Q4. How to identify suitable research gaps and expected trends in the domain?

2 Data Sources and Methodology

The systematic literature review includes a methodical and equitable search. This allows extensive and integrated assessment and analysis. The literature management is carried out using relevant bibliometric software as it permits the collection of bibliographical records and creates an outline of field knowledge. The field knowledge encompasses some scientific domains. The data is represented in form of visually enticing graphics for different knowledge spheres. Domain analysis has become a popular field that facilitates discovering large information and track development in the sphere. Citespace is a comprehensive software that enables keyword burst analysis, and timeline review providing a more complex interface [7]. Therefore, Citespace has been chosen for analysis due to its inclusiveness, consistency, and convenience of use. Following guidelines laid in PRISMA2020 [8], the data was selected from relevant sources, and it was further retrieved using screening techniques and cleaned to eliminate bias. The abstract and interactive analysis was conducted by CiteSpace. The visual representation in form of graphs in CiteSpace included nodes coloured in shades of grey to red representing evolution with time. The frequency of the label was highlighted by its size. The width of connection among nodes facilitated recognition of other parameters like network centrality, size of the cluster, silhouette, and time-dependent analysis in regards to the significance of the information. The quantitative analysis was followed by a qualitative analysis.

2.1 Data Sources

2.1.1 Description of Retrieved Keywords

The keywords “bond strength”, “repair”, “masonry”, “adhesion”, and “substrate” were used to identify the 15 most appropriate works. The thorough review of this literature aided in building a proper logic grid. Therefore, the keywords for “substrate” are “interface”, and “substrate”. Keywords for bond strength are represented as “adhesion”, “grout”, and “tensile strength of bond”.

2.1.2 Review Strategy

The first thing in this study was to identify some critical terms for bond strength investigation of substrate and repair material for masonry structures. Next, the keywords were connected using Boolean operations utilizing Web of Science (WOS) and Scopus as the literature source databases as follows: TS= ((Repair OR Adhesion OR grout) and Bond and interface and masonry not (”stone”). The search period, which covered 20 years from 2002 to 2022, was established from January 1 to December 31, 2022. This method led to the retrieval of 1430 publications from SCOPUS and 1592 papers from WOS [7]. The following actions were done to make sure the data were accurate: Choosing ”Article” as the filter for the type of literature, choosing ”English” as the language of the literature, and deciding on the scope. The construction building technology, material science multidisciplinary, civil engineering, and material science composite were selected as the focus of the literature study. The literature was manually reviewed to weed out irrelevant articles; duplicate articles were removed using CiteSpace [7]. The studies selected for each process are shown in Figure 2, and a total of 678 studies were deemed to be pertinent to the analysis.

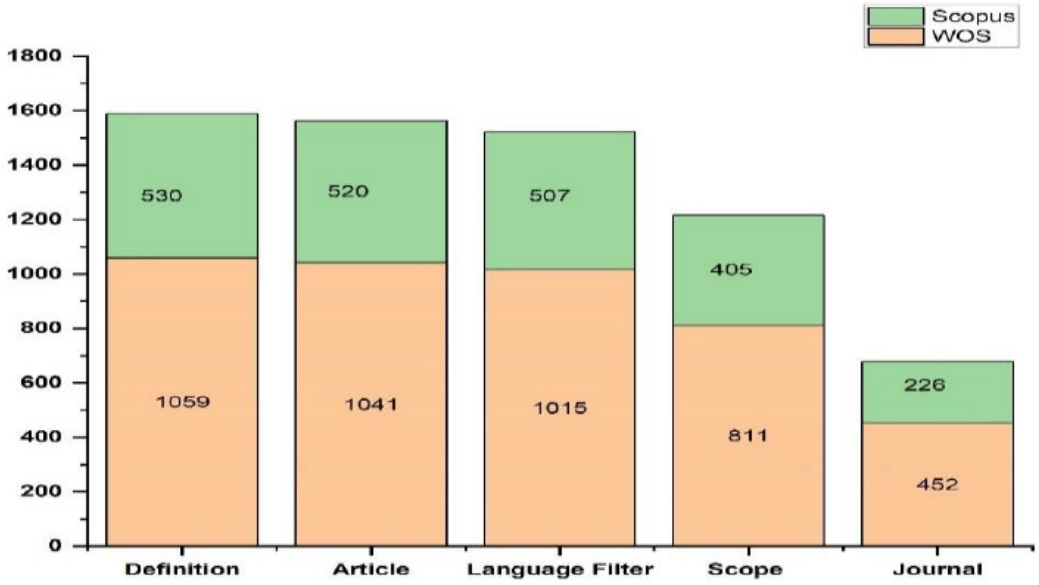


Figure 2: Data Cleaning process for articles collected from WOS and Scopus

3 Visualization Results and Analysis

3.1 Number of Published Papers

Figure 3 illustrates the statistical analysis of the literature retrieval data, which revealed that from 2003 to 2008, there were less than 20 publications. However, from 2009 to 2021, the number of articles considerably rose. The number of publications stayed around 50 over those years. Since the awareness about intervention before applying repair techniques has been enhanced over the years, an increasing number of scholars have concentrated on this topic. 2021 had a significant rise in the toll of published articles, reaching 96. From the trendline, the overall trend is in the forward direction.

3.2 Cooperative Network Analysis for Countries

Each node size represents the publications count in that country for the cooperative network among the numerous nations. According to the years 2003–2022, the node shades were coloured ranging from grey to red, with the breadth of each colour denoting the yearly number of publications. According to Figure 4, Italy had the most publications i.e., 202 followed by China, England, USA, and India. It has been observed that there is no centrality for any nation indicating independent research by different countries.

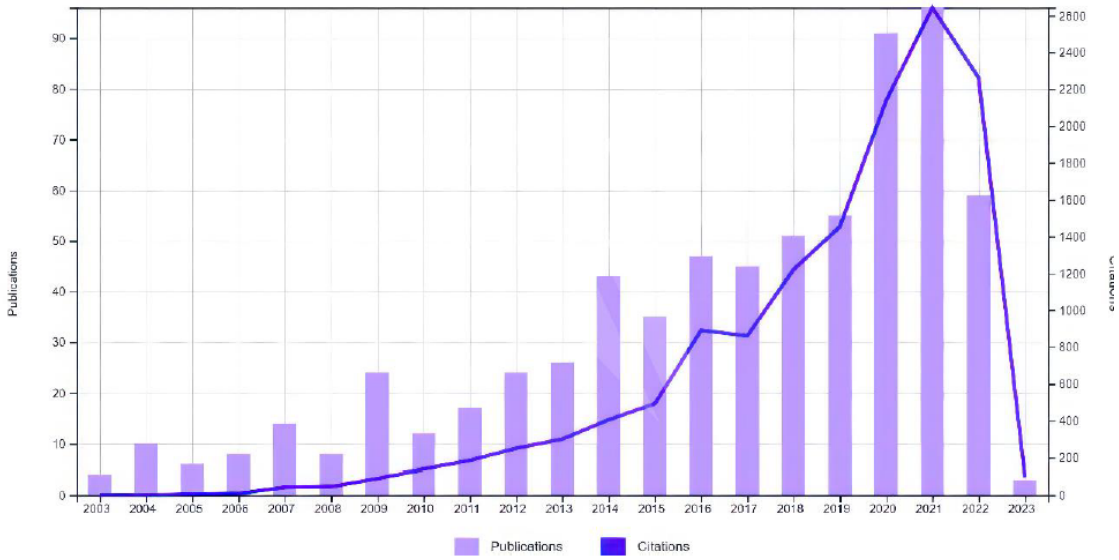


Figure 3: Number of publications and Citations based on bond strength investigation.

Countries	Number of Articles
202	Italy
53	England
53	People R China
49	USA
47	India
45	Portugal
40	Australia
31	Spain
24	Germany
22	Canada

Figure 4: The collaborative countries network

3.3 Cooperative Network Analysis for Authors and Institutions

Though, as seen in Figure 5, there were a maximum of 22 publications published simultaneously by a single author there were also a few instances of many writers working together. This demonstrated that there is not a high concentration of writers focused on this objective. Even though there was limited cooperation among institutions, the research authors were more centred and molded a core research group in their institutions, as shown by the network Figure 6 of institutional cooperation, which combined the units of authors and research institutions with the maximum publications.

3.4 Co-Citation Analysis

3.4.1 Cited Procedure

The study of the history and future directions of research on "bond strength between substrate and repair material for masonry structures" requires analyzing highly referenced literature. The period was "1 year," the node type was "Reference," and the Top 10% of each slice, up to 100, was the criteria, and Pathfinder was the network crop. The "Pathfinder" network culling technique was utilized, and the tagged terms were extracted from the pertinent citation literature using the Log-Likelihood Ratio (LLR) algorithm. According to CiteSpace's study, were referenced in three publications with the most co-citations and significantly influenced the field. Lourenco et al. suggested a constitutive model for masonry structure analysis. The discretized model is prepared to understand the slipping, cracking and crushing behaviour of the material. Valluzi et al. proposed to characterize the bonding capability of different masonry components strengthened by composites on basis of their shear action while Papanicolaou et al. developed a new material as textile-reinforced mortar to strengthen the unreinforced masonry walls.

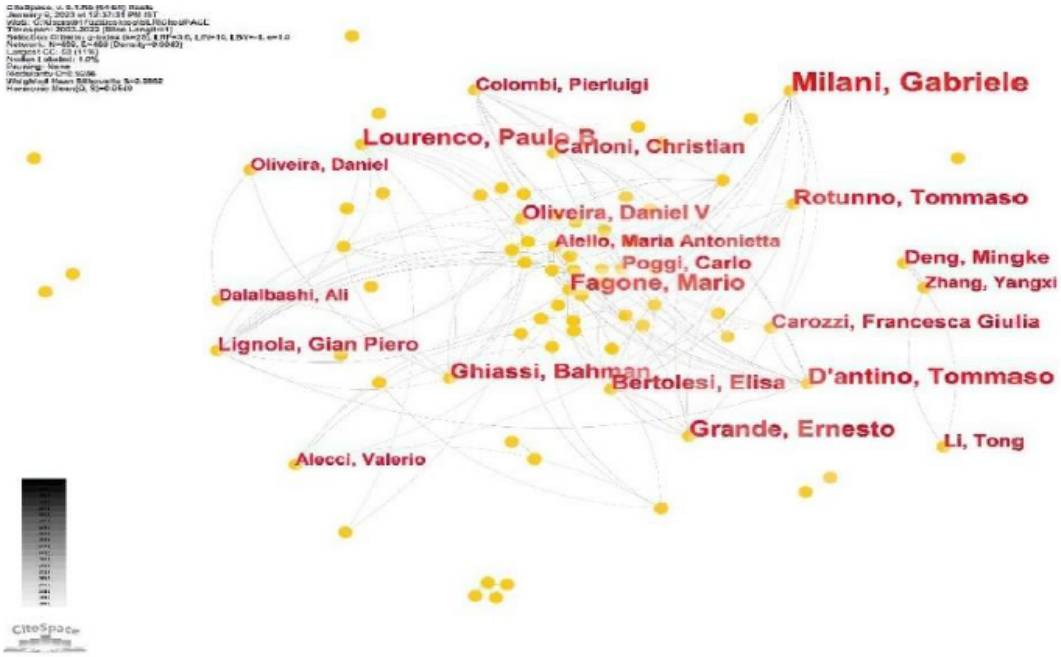


Figure 5: The collaborative authors network

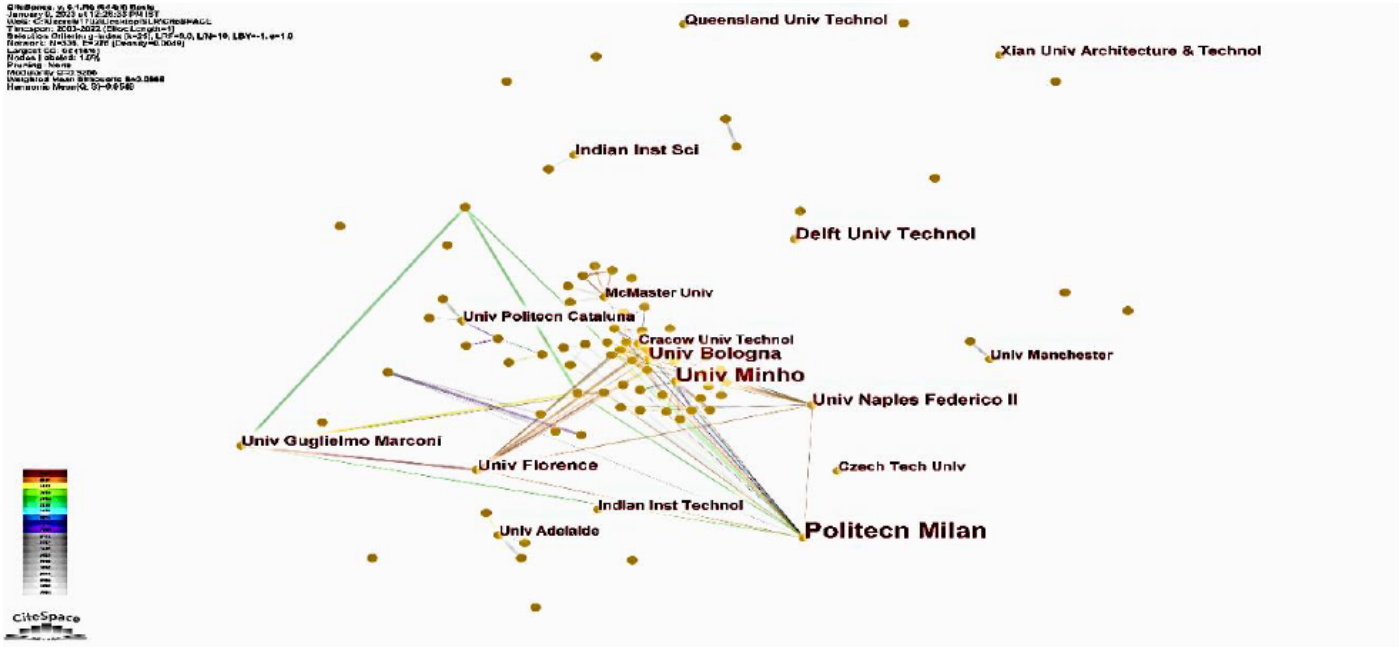


Figure 6: The collaborative institutions network working in the domain.

Table 1 highlighted the precise details of each cluster for the co-citation network. All of the clusters' silhouette scores were very near 1, indicating that there was substantial homogeneity within each cluster. In other words, the clustering results' relevance and reliability have been proven. To characterize the clusters, labels were applied using the LLR method. The biggest size of references is seen in Clusters 0 and 1. With references emerging earlier and still being continuously acknowledged in the literature, Cluster 0's average age of the literature is 2015 and Cluster 1 is 2005 indicating new material as a solution to repair on the rise.

3.5 Co-Vision Network Analysis

3.5.1 Keywords Co-Vision Network Analysis

Keywords that describe the main ideas behind current research in a discipline and indicate its fundamental substance. Given that the node type was "term," noun phrases were retrieved based on the terms in the title, abstract, authors, and publishers. The period was "1 year," and the standard was "Top 10% per slice, up to 100." Pathfinder was the web culling technique used. Figure 7 illustrates the analysis of relationship lines and keyword nodes from the database. The top 20 keywords and their occurrence in the network are detailed in Table 2.

Table 1: The co-citation Clusters summary

Cluster Label	Cluster-ID	Size	Average Year	Silhouette	Alternate Label
Tensile behaviour	0	44	2015	0.861	FRCM composite, strengthening effect
Numerical investigation	1	29	2005	0.992	Numerical investigation, brick type
HDC-masonry composite element	2	23	2014	0.997	HDC-masonry composite element, masonry cylinder
Curved masonry substrate	3	21	2015	0.905	Curved Masonry substrate, jute textile-reinforced mortar
Short PVA fibre	4	15	2017	0.906	Analytical investigation, brick masonry panel
fabric-reinforced cementitious matrix	5	14	2016	0.91	Bond behaviour, brick masonry
interlocking masonry	6	12	2014	0.978	interlocking masonry Evaluation
FRP delamination	7	9	2015	0.968	curve masonry substrate, optical sensor

Table 2: Top Keywords with their number and begin year

Keyword	Number	Begin Year	Keyword	Number	Begin Year
Behaviour	158	2010	Compressive strength	42	2003
Concrete	99	2006	Mechanical property	41	2015
Strength	78	2007	System	40	2015
Wall	78	2008	Brick masonry	37	2004
Performance	68	2010	Mortar	33	2004
Model	59	2009	Masonry wall	28	2011

In the last decade, keywords like behaviour, performance, and model have been frequently used indicating research carried out as an experimental investigation to understand the nature of materials. Concrete has been the second most used term indicating there is comparative analysis in terms of test procedures, and reference models of masonry with concrete. To test the mechanical properties and the base of different tests to determine bond strength by compressive strength justify their mentions in the literature since 2003.

To carry out the cluster analysis of the keywords in the research domain, Citespace was chosen, and the Log Likelihood Rate (LLR) algorithm was used. Based on the feature words accepted by LLR as names of the clusters, major 16 clusters were recognized. Their graph on the timeline with the generated keywords is presented in Figure 8, while Table 3 presents details of the 16 clusters corresponding to their size, label, mean year of occurrence, and a few alternate labels. The detailed information for clusters is presented below:

- Cluster #0 (frcm composite) is the cluster with the second most produced literature i.e., with 36 articles. This cluster started in 2015. The most cited terms of the cluster include frp (56), composite (51), and system (37). The most cited article is "Freeze-thaw effects on the performance of trm-strengthened masonry". The author proposed an analytical and experimental exploration of the durability of masonry structures reinforced by TRMs [9]. Though the composite behaviour of the material showed deterioration, the freeze-thaw cycle did not affect the performance of the panels. Researchers are inculcating the use of different composites to strengthen old structures; few authors proposed the use of a fibre-reinforced matrix to enhance the mechanical properties of the structures. Most studies include testing tensile strength and conducting the shear test to understand failure behaviour [10–12].

- Cluster #1 (Rendering mortar) is the largest cluster with 40 articles. The cluster began in 2015. The cluster mainly discusses the mechanical properties and compressive strength with its mention 40 and 41 times respectively in literature as the keyword. The cluster highlights the different case studies related to enhancement made in the mortars attributing to the "Reinforcing rammed earth with plant fibres: A case study" article by Koutous A [13]. The authors investigated the performance of rammed earth with additives like cement, barley, and date palms in terms of compression and tensile characteristics. This was concluded that though natural fibres reduced stiffness they overall improved the properties of the earth[14].

- Cluster #2 (Debonding Resistance) is the third cluster with 31 articles. This cluster mainly highlighted the response of the unreinforced masonry components, primarily walls. The different failure mechanisms and fracture patterns are reported by the researchers. They determined the effect of multilayered loading on the URM wall on basis of the macro model and proposed interaction curves [15]. This cluster proposed numerical models to investigate the behaviour of the masonry subjected to yield forces [16].

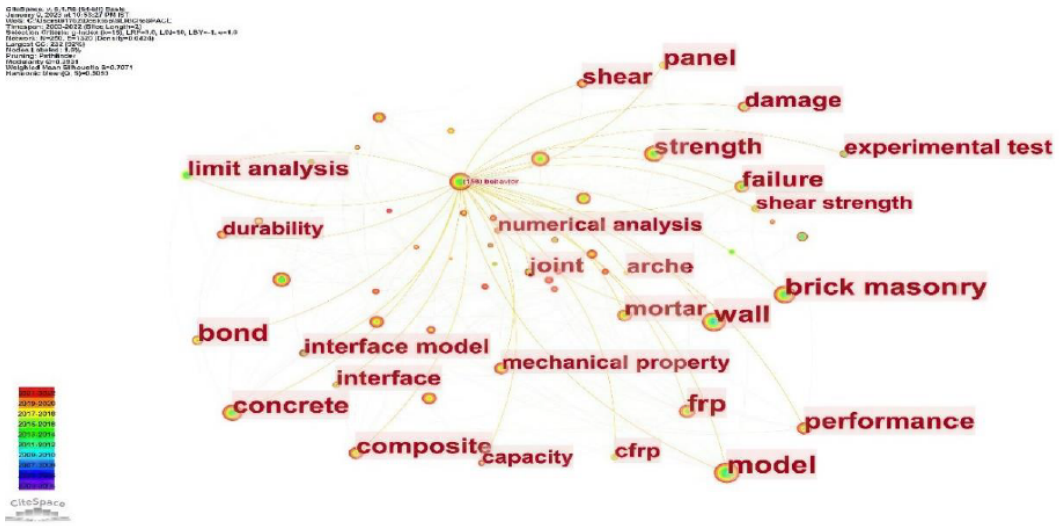


Figure 7: The Keyword Network for bond strength investigation of the substrate with repair material

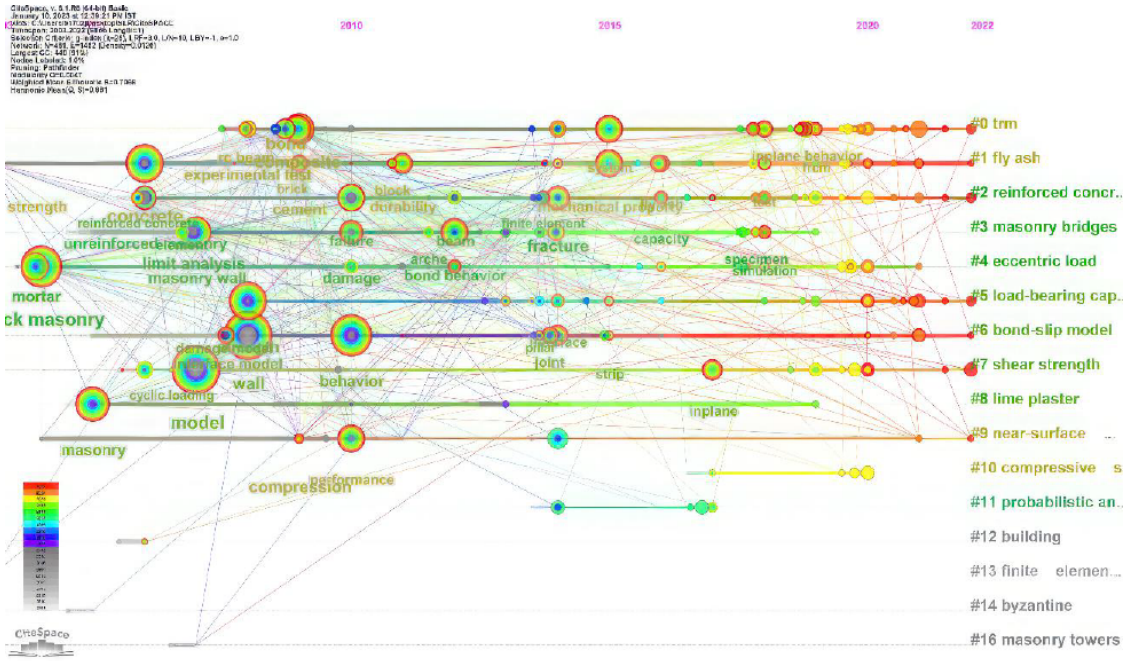


Figure 8: The cluster network of keywords addressing bond strength of materials.

- Cluster #3 (Masonry bridges) also emphasized the preparation of numerical models to examine the effect of repair materials on the components of structures. The authors applied carbon fibre-reinforced polymer reinforcement on the pillars of the masonry bridge to determine debonding mechanism in terms of ductility and friction-cohesive relationship [17].
- Cluster #4 (masonry column) included discussions about brick masonry primarily. The robustness response of the structures when subjected to different environmental impacts is highlighted. Researchers investigated the impact of efflorescence on the structure. They created artificial conditions for the development of efflorescence and subsequently proposed solutions for it [18]. They identified ettringite carbonation as the cause of efflorescence and also discussed its genetics.
- Cluster #5 (composite assemblies) was more closely related to the bond behaviour. It mainly includes "strength", "interface", and "bond strength" as the cited members of the cluster.
- Cluster #6 (a masonry structure) also constitutes articles that investigate the fracture behaviour of reinforcement to enhance the strength of the masonry structures [19]. This cluster also suggests analyzing the micromechanical behaviour of the structure as joints, cracks, and voids can complicate its study [20, 21].
- Cluster #7 (concrete masonry wall) is also composed of articles proposing the interlocking nature of materials [22, 23].

- Cluster #8 (weathering resistance), Cluster #10 (high temperature) and Cluster #14 (characterization) are related to the durability aspect of the structure. They highlight using various experimental techniques to investigate the physical, chemical, mineralogical, and thermal characteristics of the materials. As the practices to restore and repair historical structures are on the rise, many case studies are conducted to individually determine the properties of the constituent materials for various structures [24–26]. Authors used techniques like XRD, FTIR, and SEM to analyze the historical lime at Hampi, India and showed the presence of aragonite and vaterite as polymorphs of calcite in lime. Similarly, Hornikova et al. carried out a study for the physical, thermal, hygral, and mechanical properties of cementitious composites using experimental setups for elevated temperature conditions. They observed that elevated temperature compressive strength remained constant only the gas permeability enhanced for all the composites [27].

- Cluster #11 (sensitivity analysis) suggests numerical techniques to understand the effect of material parameters. Researchers prepared a micromechanical Finite element model for concrete masonry and used simulations like Monte Carlo to perform sensitivity analysis for the assessment of parameter effects [28].

- Cluster #12 (dynamic investigation) also underlined the application of numerical methods for parameter identification.

- Cluster #13 (Quality Control) suggests improvised techniques like infrared imaging for characterization. Starnes et al. proposed a study of the woven fabrics embedded in concrete to enhance their strength with the use of infrared thermography [29]. He employed both controlled-flaw experiments and FE techniques and was able to derive good agreement in the results. This gives a new dimension to the different characterization techniques for structures to understand their constituent properties [30]. The detailed information for the clusters identified in the keyword network, including their size, silhouette scores, alternate labels, and average years of occurrence, are summarized in Table 3. This table provides a comprehensive overview of the 16 major clusters recognized in the study, offering insights into the key themes and trends in the research domain.

3.5.2 Analysis based on Keyword Burst

This investigation used a burst detection of keywords using an algorithm built into the software to better classify research areas in the subject, and the findings are displayed in Figure 9. There were 10 burst keywords discovered, along with their burst strength, start year, end year, and duration. The longest-lasting term was "limit analysis," which persisted for six years from 2010 to 2016. This was followed by the term "failure" and "masonry wall" which lasted three years, from 2010 to 2013 and 2011 to 2014 respectively. "Concrete and masonry structure" is the most powerful term. This shows that most of the investigation concerning masonry structures is conducted about concrete behaviour.

Top 10 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2003 - 2022
limit analysis	2007	4.44	**2010**	2016	
failure	2010	3.93	**2010**	2013	
masonry wall	2007	3.43	**2011**	2014	
strength	2008	3.2	**2012**	2014	
concrete and masonry structure	2014	7.42	**2014**	2017	
concrete structure	2014	3.17	**2014**	2016	
fly ash	2016	3.44	**2016**	2018	
performance	2010	4.79	**2019**	2022	
trm	2019	3.92	**2019**	2020	
analytical model	2019	3.48	**2019**	2020	

Figure 9: The Keyword with Citations Burst in the articles related to the domain

4 Systematic Review and Summary of Literature

On the basis of the bibliometric analysis presented in the "Visualization results and analysis" segment, a macro-level outline of the existing research in the field is shown, providing an inclusive understanding of the existing state of advance and future aspects of the complete field. The bibliometric analysis technique is still insufficient to provide a tangible overview of the inherent significance of the field. To address this limitation, the present study conducted a content review of the bond between the surface of historical structures with the applied repair material.

Table 3: Clusters for keyword network

Cluster-ID	Cluster Name	Size	Silhouette	Alternate Label	Average Year
1	Rendering mortar	40	0.735	Mechanical properties, gypsum efflorescence	2015
0	frcm composite	36	0.671	frcm composite, ambient vibration testing	2015
2	Debonding resistance	31	0.757	Unreinforced masonry wall, masonry component	2016
3	Masonry bridge	27	0.866	Masonry bridge, CFRP plate	2013
5	Composite assemblies	26	0.764	Mechanical properties, buckling behaviour	2016
6	Masonry structure	24	0.809	Masonry wall, bond-slip behaviour	2013
4	Masonry column	22	0.907	Masonry column, traditional masonry wall	2012
7	Concrete masonry wall	21	0.688	Masonry wall, geometric parameter	2012
9	Of-plane behaviour	13	0.825	Brick masonry, factors assessment	2014
8	Weathering resistance	11	0.925	Weathering resistance, ambient vibration testing	2013
11	Sensitivity analysis	7	0.956	Sensitivity analysis, ambient vibration testing	2016
10	High temperature	5	0.962	High temperature, masonry wall	2019
12	Dynamic investigation	2	0.997	Masonry bell tower, masonry wall	2006
13	Quality control	2	1	Preliminary thermography studies for quality control of concrete structures strengthened with fibre-reinforced polymer composites	2003
14	Characterization	2	0.997	Investigative methods for the characterization of historic mortars - part 1: mineralogical characterization	2005
16	Ambient vibration testing	2	0.993	Ambient vibration testing of historic masonry towers for structural identification and damage assessment	2007

4.1 Mortar in historical structures

One of the most prominent historical materials is masonry. It is a traditional, extensively used, enormously supple, and efficient construction material. It has considerable potential for future development owing to its sturdiness, durability, and appealing reasons. It is an assorted material with components of relatively indefinite geometry and a high mechanical inconsistency. This makes characterizing its constituent materials a subject of great interest with respect to the appropriate intervention techniques. Furthermore, the proper compatibility assessment of the materials is very essential to restore the structure as the bond between composite materials forms the weakest bond [31]. Mortars with diverse types of binders are used in the primeval period for many applications. Bricks or stones bonded with masonry mortar, wall finishing by plaster as well as render, mortar for floor finishing, infill walls comprising of rubble mortar, water conduits casing, etc. This has been observed that there exists great composition disparity in different mortars owing to their geographical location and period of procuring [32]. For about two centuries mud, gypsum, and lime have been the most used binder material, however, gradually their use was reduced with the introduction of natural cement followed by Portland cement which dominates the construction market today. The most ancient binder is mud mortar which can be identified with the use of clay in 6000 BC evident at Catal Huyuk in Turkey. A terrazzo floor at Canjenü in Eastern Turkey laid with lime mortar dates back to 5000 BC. Gypsum is the most used binder in Pharaonic Egypt, in the middle east, Germany, and Paris during the medieval period. In Europe, studies by Adams et al. showed that most Cathedrals near Paris had gypsum as a binder. Though gypsum and mortars are also prominent historical binders research shows that lime mortar supersedes them globally during the period [33]. Previously, the historic mortar characterization was done using traditional wet chemical analysis.

This method provides the result with cumbersome interpretations due to limited knowledge about the characteristics of the constituents of the mortar. However, the present mortar characterization techniques suggest optical microscopy and X-Ray diffraction methods as the initial step for qualitative determination of constituents of the mortar. These methods include analysis techniques namely SEM, FTIR, TGA, etc. [8, 34, 13] The appropriate method depends on the requirement and material availability. The selection also depends on the required intervention for the conservation of ancient structures the requirement is information about the composition of the compatible repair material, causes of degradation of old structures, and the distinction between various building phases with time [15]. The basic parameters that need to be identified for finding the most compatible repair material are the hydraulicity of the binder and the proportion and grading of the aggregate corresponding to the binder and constituents of the mortar [17]. However, in aspects of archaeology, the determining of chronology, and spatial distribution play significant roles to determine socio-economic deductions about raw material and its production. The recent development in the field of characterization involves details about the technique of procuring and processing material by burning, mixing, hydration, and carbonation and identifying mineral composition. Inclusive information about lime and old mortars has been at the interest of researchers for the last decade [4].

4.2 Repair material for historical structures

The major requirement for restoration work is to safeguard original renders and to fill repair the missing part of the structure with suitable mortar. The selection of the appropriate material is critical as if a material like cement is chosen, being strong and impermeable it would further accelerate the deterioration of the structure [35] [18]. The other important factor is durability as the structure is exposed to adverse weather conditions, moisture retention, and salt accumulation. Studies show that durability outweighs other parameters in the case of ancient structures. Therefore, a judicious decision needs to be made while selecting repair material [36].

Mortars are composed of binders, aggregates, and water with required additives. Therefore, varying the compositions of these parameters facilitates achieving the desired mechanical and physical properties. Binders are solely responsible for influencing the properties of fresh as well as hardened mortar [37]. The different mortars used as repair material include Non-hydraulic Lime, which improves the workability of existing mortars with hydraulic binders; Hydraulic lime, which imparts higher strength; Pozzolana lime binders, possessing properties similar to hydraulic lime and most historically used binder; calcium silicate cement, possessing shorter setting time with higher strength; and clay earth mortars, empirically suggesting longer service life. The mortar for repair action depends on the function of the mortar and inhibited masonry typology itself. As the conditions keep changing, it is possible that the intervention with a new repair mortar could be for conditions different from the situation encountered by the mortars used originally [29, 27]. The different requirements to be met for selecting repair material are discussed in detail in Table 4.

4.3 Bond of Materials with Repairs at Interface

The adhesive bond formation is the basic necessity to achieve proper intact between the substrate of the structure and the repair material. Many theories highlight the mechanism of adhesive formation [38]. The first theory is based on mechanical interlocking that occurs with the suitable adhesive material penetrating through the pores on the specimen surface [6]. The second theory takes into account the chemical bonds on the interface. The contact at phases determines bond strength [39]. The transition theory elucidates the formation of the surface at repair material and substrate contact. The adsorption theory states that the adhesive material used for repairs adheres to the wetted surface by interatomic forces [40, 41]. This theory also comprehends the notions of rheology, surface energy, electrostatic and diffusion concepts. It is very difficult to discern the most suitable theory but it can be inferred that the substrate of the structure and repair material bond is majorly dependent on the adhesive capability of the interface, interface roughness by the interlocking of aggregate and many time-varied conditions [24, 25]. Further, these characteristics depend on factors like adhesion governed by the bonding agent, compaction of the sample, roughness, age, and dampness of the repair material surface. The size, shape, and surface preparation of the aggregate influences the interlocking of the aggregate and its friction [26, 42].

There are numerous investigations conducted employing standard tests to measure bond strength [9]. The test is chosen to correspond to the state of stress and required applications like slant shear test for epoxy and latex bonding agents for compression and shear combined state. One of the most important aspects of repairing an old structure is the bond of the material with the substrate. This is because it forms mechanically the weakest linkage between the structure and binding agent [30]. It severely affects the resistance and durability of the masonry. Many studies have reported that the adhesion of the binding agent is more significant than any other mechanical property of the formulation [43]. It has also been concluded that the bond between the mortar joint and unit controls the compressive strength of the structural element rather than the compressive strength of each element individually. The cases reported separation of outer and inner walls owing to weak bond strength [31]. The most prominent seismic failures, i.e., out-of-plane failure causing wall collapse, are directly connected to bond strength. Therefore, this necessitates the need to understand the adhesive tendency of the mortars used in restoration and repair work [32]. The main objective of the mortar in repair work is to connect various components of masonry. To achieve this the tensile strength and adhesive capacity of the material should be focused on rather than other properties. This has also been concluded that the behaviour of the bond does not directly depend on the mechanical characteristics of the binding material [33].

Table 4: Aspects for Selection of Repair Material [28]

Design Aspects for Selection of Repair Material	Details
Functional Requirements	
Conservative Issues	Historical authenticity, Reversibility i.e., removal of old mortar without damage to the structure
Aesthetic Issues	Visual appearance, Match original material in color, texture, etc., Provide a surface for decoration
Structural Issues	The efficiency of flexibility and strength to bear loads
Service Life Issues	Resistance to expected environmental loads, No negative impact on the existing structure, Provide coating to protect underlying masonry
Technical Requirements	
Requirements for renders	Since renders are used in exterior applications, therefore, their service life is impacted by environmental exposure
Requirements for plasters	Plasters are used for interior applications therefore; strength plays a dominant role.
Techniques of application	Suitable consistency of the mortars, Deterrence of premature setting, Measures for elevated temperature, sturdy wind, and relative humidity, The predetermination of regime and curing period
Performance Requirements	
Common Requirements	Not to damage the existing substrate, possess higher strength, Be Flexible and strong in adhesion, low tendency to shrinkage and resistance to environmental issues
Specific Requirements for Renders	Modest water absorption and capacity of drying and possess surface strength, Resistance to soluble, releasable salts and freeze-thaw cycles
Specific Requirements for Plasters	Resistance to soluble, releasable salts and freeze-thaw cycles

The grouting material may be weak in mechanical properties but can be effective and durable. Therefore, bond strength knowledge is very essential for selecting appropriate repair solutions [44–46].The major gaps identified in the study are discussed as follows:

- **Complex Nature of Ancient Mortar:** There are significant disparities in the constituents of ancient mortar [10, 11]. Ancient masonry is intricate, mostly unreinforced or with traditional reinforcement, and often integrated into complex structures. This complexity makes it difficult to separate architectural and structural features.
- **Variations in Modern Masonry:** Modern masonry exhibits substantial disparities in materials and building technology. Traditional and local technologies vary from one country to another [47]. This diversity leads to a lack of awareness and understanding, posing barriers to the proper rehabilitation of structures.
- **Understanding Bond Parameters:** There is a lack of understanding of the significant parameters affecting the bond between mortars and units [12]. Factors such as differential volume changes between the unit and mortar, which can lead to water penetration and separation, are often overlooked. The importance of the durability of bonds is frequently neglected in practice [14]. Similarly, shrinkage, which may reduce the life expectancy of the bond, is also a neglected parameter. There is a need for more research to recognize factors influencing mortar and unit bonds, as most standards and tests are limited to the strength of the bond [48].
- **Missing Documentation:** Significant documents and drawings detailing the renderings used in old structures are often missing [16]. The inadequate intervention in ancient structures, combined with a lack of resources for modern techniques, has led to the loss of novel finishings and rendering [49]. This has created issues related to the incompatibility of old materials with new repair materials due to variations in mechanical and physical characteristics, such as strength, porosity, color, and chemical composition [22, 50].

5 Outlook

Structural strengthening for historical masonry structures using various repair composites has become a subject of interest nowadays, offering a cost-effective solution globally. The systematic Literature Review identified and examined many research barriers, which this study sought to address. Additionally, as shown in Table 4, the 14 clusters may be categorized into each phase of the conceptual model for bond strength investigation for substrate and repair material for masonry structures. This conceptual model is combined with four major research gaps identified in the literature cluster, indicating a significant demand for literature to bridge these gaps and remove obstacles. The first area of weakness is "Limited Finite Element Modeling"; there are three clusters connected to it. Proper numerical investigations aiming at assessing the debonding phenomenon of materials used as reinforcements to masonry structures are needed.

The second gap is the lack of required "understanding of mechanical attributes", highlighted by three clusters. Appropriate knowledge of the bond behaviour of different retrofitting solutions when applied on masonry substrates is crucial. There is a need for proper setups for bond tests, analyzing results in terms of failure modes, curves, strain profiles, and interface laws. "Developing numerical models" is essential to understand the behaviour of loads on the structure and interpret fracture and crack patterns. Four clusters majorly emphasize this aspect. Many researchers are preparing numerical models to validate their experimental investigations related to the bond strength of substrates with repair alternatives. FRP composites are becoming popular as repair composites for masonry structures. Researchers are also trying to understand the performance of these alternatives in brick masonry by employing macro modelling as a finite element technique. The fourth research gap is the demand for the expansion of characterization practices. Four clusters suggest carrying out characterization studies to guide the selection of suitable repair material corresponding to the structural requirements. This study extracted example publications from each literature cluster and compiled data to better identify specific research needs. The overview of the study is presented in Figure 10.

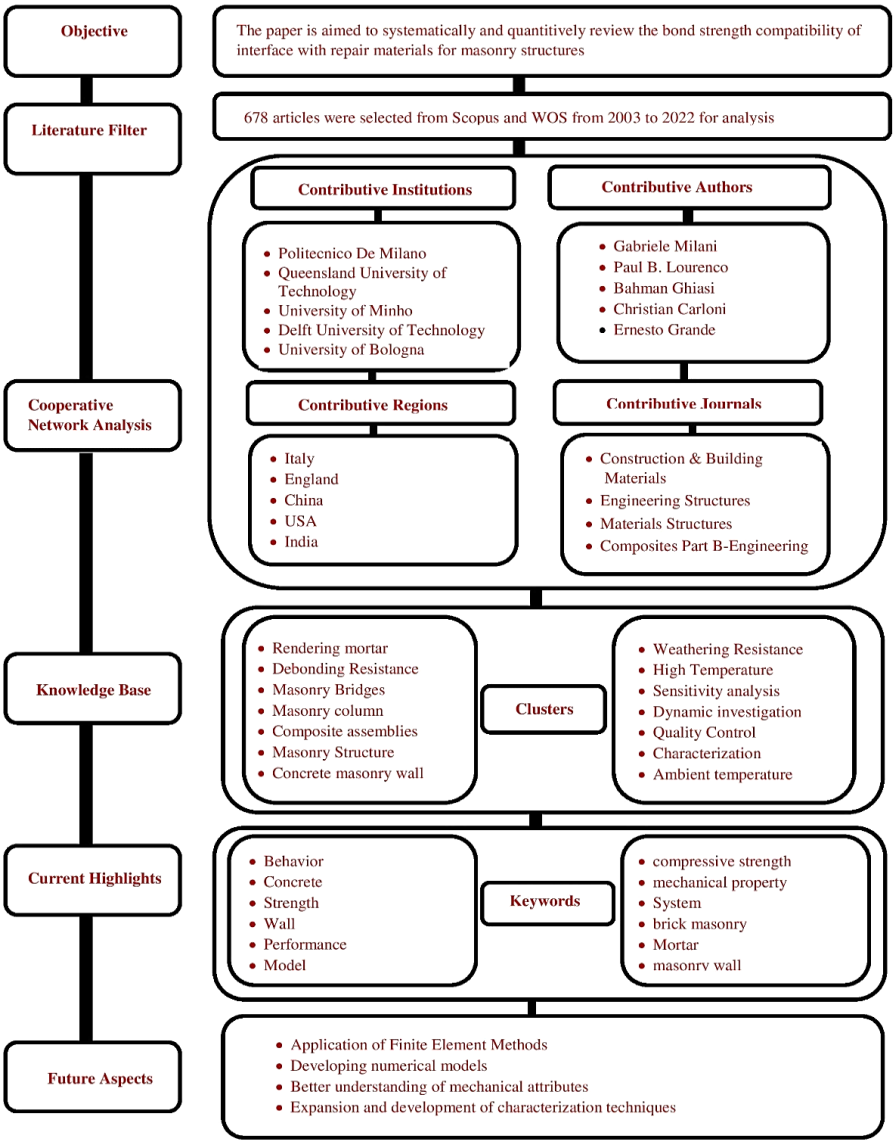


Figure 10: The overview of the systematic literature review.

6 Conclusion

In this study, 678 articles on bond strength investigation of substrate and repair material for masonry structures were analyzed and elaborated using CiteSpace. These articles included published countries, authors, authors' institutions, and papers, keywords, and clusters of keywords. CiteSpace was used to analyze and elaborate on these articles from 2003 to 2022. As a result, this study may highlight areas in need of improvement for investigating the bond strength of the original material with repair materials and offer recommendations for further investigation. It was evident from the bibliometric analysis of the articles in the literature studied that Italy had the most publications worldwide, followed by China, the US, England, and India. This reflects the growing awareness of the practices to retain old structures as epitomes of culture and heritage, with many nations contributing significantly to the domain. Additionally, researchers are exploring alternatives to old renders by characterizing them. According to the contributions and impact of the lead authors and institutions discovered in the analysis network, Gabriele Milano, from Politecnico Di Milano, is the most productive author in the area of masonry materials and its analysis. Three of the top ten publishing institutes are Politecnico Di Milano, the University of Minho, and the University of Bologna. The total collaboration network and publishing output, however, have a lot of space for growth and improvement.

In the present study, the top 10% of co-cited references were analyzed in a time slice that resulted in 527 nodes and 2712 link lines. Most of the references cited majorly proposed preparing a constitutive model to carry out structural analysis. Eight major clusters were recognized based on co-cited references, covering aspects from tensile behaviour and frcm composite to strengthening behaviour, numerical modelling, and analysis of different composite materials. Recent keywords include "limit analysis," "failure," "masonry wall," "strength," and "performance," indicating areas for ongoing academic examination. However, there is a need for further research on terms related to the bonding between different materials and the characterization of old mortars. Future research should be aligned with carrying out Finite Element Modelling for existing masonry structures to analyze and simulate their behaviour, determining the mechanical attributes of materials with various laboratory tests including SEM, XRD, FTIR, TGA, etc., preparing numerical models to predict the bonding tendency for different materials, and characterizing the constituents of the original as well as repair material. The extensive content review suggests that though masonry structures have served us for years, they are deteriorating due to weather conditions and age. The lack of proper documents from previous eras leaves us deprived of original drawings and the proper composition of constituent materials. Therefore, many studies are being carried out to implement modern techniques and devices to monitor the health of structures and detail their composition. The disparity in properties between present and ancient materials affects the practice of repair and rehabilitation. Thus, numerous investigations suggest preparing composite materials that enhance structural properties without compromising aesthetics.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

References

[1] A. Moropoulou and A. Bakolas, "Criteria and methodology for restoration mortars compatible to the historic materials and structures," 2000. Accessed: Mar. 01, 2023.

[2] "The world heritage fund / le fonds du patrimoine mondial statement of assessed contributions as at 30 november 2022 / etat des contributions mises en recouvrement au 30 novembre 2022."

[3] G. Ma, L. Wang, and Y. Ju, "State-of-the-art of 3d printing technology of cementitious material—an emerging technique for construction," *Sci China Technol Sci*, vol. 61, no. 4, pp. 475–495, 2018.

[4] M. E. Loke, K. Pallav, and R. Haldenwang, "Characterisation of heritage cementing materials for restoration purposes: A review," *Journal of the South African Institution of Civil Engineering*, vol. 62, no. 1, pp. 10–21, 2020.

[5] D. Y. Wu, S. Meure, and D. Solomon, "Self-healing polymeric materials: A review of recent developments," *Progress in Polymer Science*, vol. 33, no. 5, pp. 479–522, 2008.

[6] C. Ince, M. A. Carter, and M. A. Wilson, "The water retaining characteristics of lime mortar," *Materials and Structures*, vol. 48, no. 4, pp. 1177–1185, 2015.

[7] S. Feng, H. Xiao, and H. Li, "Comparative studies of the effect of ultrahigh-performance concrete and normal concrete as repair materials on interfacial bond properties and microstructure," *Engineering Structures*, vol. 222, p. 111122, 2020.

[8] I. D. Gavardinas, A. Athanasoulas, K. Spanos, A. D. Giannoukas, and A. E. Giannakopoulos, "Novel methods for the mechanical characterization of patches used in carotid artery repair," *Materials Science and Engineering C-Materials for Biological Applications*, vol. 93, pp. 640–648, 2018.

[9] A. Dalalbashi, B. Ghiassi, and D. v. Oliveira, "Freeze-thaw effects on the performance of trm-strengthened masonry," *Journal of Building Engineering*, vol. 59, p. 105077, 2022.

[10] F. Bompadre and J. Donnini, "Fabric-reinforced cementitious matrix (frcm) carbon yarns with different surface treatments embedded in a cementitious mortar: Mechanical and durability studies," *Materials*, vol. 15, no. 11, p. 3927, 2022.

[11] A. Khaloo, H. Moradi, A. Kazemian, and M. Shekarchi, "Experimental investigation on the behavior of rc arches strengthened by gfrp composites," *Construction and Building Materials*, vol. 235, p. 117519, 2020.

[12] A. N. Al-Gemeel, Y. Zhuge, and O. Youssf, "Experimental investigation of basalt textile reinforced engineered cementitious composite under apparent hoop tensile loading," *Journal of Building Engineering*, vol. 23, pp. 270–279, 2019.

[13] S. Feng, H. Xiao, R. Zhang, and C. Yang, "Bond performance between substrate concrete and repair mortar: Effect of carbon fibre and expansive agent," *Construction and Building Materials*, vol. 250, 2020.

[14] A. Koutous and E. Hilali, "Reinforcing rammed earth with plant fibers: A case study," *Case Studies in Construction Materials*, vol. 14, p. e00514, 2021.

[15] N. Bianco, A. Calia, G. Denotarpietro, and P. Negro, "Hydraulic mortar and problems related to the suitability for restoration," *Periodico di Mineralogia*, vol. 82, no. 3, pp. 529–542, 2013.

[16] D. G. Aggelis, "Damage characterisation of inhomogeneous materials: Experiments and numerical simulations of wave propagation," *Strain*, vol. 47, no. 6, pp. 525–533, 2011.

[17] R. Lumantarna, D. T. Biggs, and J. M. Ingham, "Compressive, flexural bond, and shear bond strengths of in situ new zealand unreinforced clay brick masonry constructed using lime mortar between the 1880s and 1940s," *Journal of Materials in Civil Engineering*, vol. 26, no. 4, pp. 559–566, 2014.

[18] R. Skominas, V. Gurskis, R. Sadzevicius, V. Damulevicius, and A. Radzevicius, "Evaluation of cement mortar suitability for repairing concrete in hydraulic structures," *KSCE Journal of Civil Engineering*, vol. 21, no. 7, pp. 2814–2820, 2017.

[19] A. Mallat and A. Alliche, "Mechanical investigation of two fiber-reinforced repair mortars and the repaired system," *Construction and Building Materials*, vol. 25, no. 4, pp. 1587–1595, 2011.

[20] R. D. F. Moreira, M. de Moura, M. A. v. Figueiredo, R. L. Fernandes, and J. P. M. Goncalves, "Characterisation of composite bonded single-strap repairs under fatigue loading," *International Journal of Mechanical Sciences*, vol. 103, pp. 22–29, 2015.

[21] J. T. Kolawole, O. B. Olalusi, and A. J. Orimogunje, "Adhesive bond potential of compressed stabilised earth brick," *Structures*, vol. 23, pp. 812–820, 2020.

[22] K. Labropoulos and A. Moropoulou, "Ground penetrating radar investigation of the bell tower of the church of the holy sepulchre," *Construction and Building Materials*, vol. 47, pp. 689–700, 2013.

[23] F. Qamar, T. Thomas, and M. Ali, "Improvement in lateral resistance of mortar-free interlocking wall with plaster having natural fibres," *Construction and Building Materials*, vol. 234, p. 117387, 2020.

[24] M. A. Salas, J. Gadea, S. Gutiérrez-González, M. Horgnies, and V. Calderón, "Recycled polyamide mortars modified with non-ionic surfactant: physical and mechanical strength after durability tests," *Materials and Structures*, vol. 49, no. 8, pp. 3385–3395, 2016.

[25] M. Asad, M. Dhanasekar, T. Zahra, and D. Thambiratnam, "Characterisation of polymer cement mortar composites containing carbon fibre or auxetic fabric overlays and inserts under flexure," *Construction and Building Materials*, vol. 224, pp. 863–879, 2019.

[26] A. Dalalbashi, B. Ghiassi, and D. v. Oliveira, "Textile-to-mortar bond behaviour in lime-based textile reinforced mortars," *Construction and Building Materials*, vol. 227, p. 116682, 2019.

[27] L. Bogoviku and D. Waldmann, "Experimental investigations on the interfacial bond strength of the adhered mortar paste to the new mortar paste in a recycled concrete matrix," *Construction and Building Materials*, vol. 347, p. 128509, 2022.

[28] L. Czarnecki, R. Geryło, and K. Kuczyński, "Concrete repair durability," *Materials*, vol. 13, no. 20, p. 4535, 2020.

[29] M. A. Starnes, N. J. Carino, and E. Kausel, "Preliminary thermography studies for quality control of concrete structures strengthened with fiber-reinforced polymer composites," *Journal of Materials in Civil Engineering*, vol. 15, no. 3, pp. 266–273, 2003.

[30] D. Cairovic, A. de la Fuente, M. Zlamal, J. Venclovsky, F. Girgle, O. Janus, and P. Stepanek, "Suitability of different tests for characterization of the dimpled concrete-to-concrete interface," *Structural Concrete*, vol. 22, no. 2, pp. 757–774, 2021.

[31] K. V. M. Rao, B. V. V. Reddy, and K. S. Jagadish, "Flexural bond strength of masonry using various blocks and mortars," *Materials and Structures*, vol. 29, no. 2, pp. 119–124, 1996.

[32] L. Toniolo, A. Paradisi, S. Goidanich, and G. Pennati, "Mechanical behaviour of lime based mortars after surface consolidation," *Construction and Building Materials*, vol. 25, no. 4, pp. 1553–1559, 2011.

[33] D. Badagliacco, B. Megna, and A. Valenza, "Induced modification of flexural toughness of natural hydraulic lime based mortars by addition of giant reed fibers," *Case Studies in Construction Materials*, vol. 13, p. e00425, 2020.

[34] B. Li and E. S. S. Lam, "Influence of interfacial characteristics on the shear bond behaviour between concrete and ferrocement," *Construction and Building Materials*, vol. 176, pp. 462–469, 2018.

[35] J. L. Parracha, J. Lima, M. T. Freire, M. Ferreira, and P. Faria, "Vernacular earthen buildings from leiria, portugal - material characterization," *International Journal of Architectural Heritage*, vol. 15, no. 9, pp. 1285–1300, 2021.

[36] K. Horníková, L. Scheinherrová, R. Štefan, and M. Foglar, "Experimental investigation of physical, thermal, hygral and mechanical properties of cementitious composites at high temperatures," *Construction and Building Materials*, vol. 255, p. 119244, 2020.

[37] H. Brocken and T. G. Nijland, "White efflorescence on brick masonry and concrete masonry blocks, with special emphasis on sulfate efflorescence on concrete blocks," *Construction and Building Materials*, vol. 18, no. 5, pp. 315–323, 2004.

[38] A. Momayez, M. R. Ehsani, A. A. Ramezanianpour, and H. Rajaie, "Comparison of methods for evaluating bond strength between concrete substrate and repair materials," *Cement and Concrete Research*, vol. 35, no. 4, pp. 748–757, 2005.

[39] G. S. Pavan, S. N. Ullas, and K. S. Nanjunda Rao, "Shear behavior of cement stabilized rammed earth assemblages," *Journal of Building Engineering*, vol. 27, p. 100966, 2020.

[40] X. Ding and Z. Yang, "Knowledge mapping of platform research: a visual analysis using vosviewer and citespace," *Electronic Commerce Research*, vol. 22, no. 3, pp. 787–809, 2022.

[41] F. Focacci and C. Carloni, "Periodic variation of the transferable load at the frp-masonry interface," *Composite Structures*, vol. 129, pp. 90–100, 2015.

[42] A. Miltiadou-Fezans and T. P. Tassios, "Penetrability of hydraulic grouts," *Materials and Structures*, vol. 46, no. 10, pp. 1653–1671, 2013.

[43] N. Bianco, A. Calia, G. Denotarpietro, and P. Negro, "Hydraulic mortar and problems related to the suitability for restoration," *Periodico di Mineralogia*, vol. 82, no. 3, pp. 529–542, 2013.

[44] H.-S. Youm, W.-Y. Lim, S.-G. Hong, and C. Joh, "Interface shear strength between ultra-high-performance concrete and normal-strength concrete," *ACI Structural Journal*, vol. 118, no. 2, pp. 287–298, 2021.

[45] G. Peng, D. Niu, X. Hu, B. Pan, and S. Zhong, "Experimental study of the interfacial bond strength between cementitious grout and normal concrete substrate," *Construction and Building Materials*, vol. 273, 2021.

[46] A. D. Espeche and J. Leon, "Estimation of bond strength envelopes for old-to-new concrete interfaces based on a cylinder splitting test," *Construction and Building Materials*, vol. 25, no. 3, pp. 1222–1235, 2011.

[47] S. Tringali, "The partial reconstruction design of the cathedral of noto," *Construction and Building Materials*, vol. 17, no. 8, pp. 595–602, 2003.

[48] L. K. Sing, N. Yahaya, A. Valipour, L. Zardasti, S. N. A. Azraai, and N. M. Noor, "Mechanical properties characterization and finite element analysis of epoxy grouts in repairing damaged pipeline," *Journal of Pressure Vessel Technology-Transactions of the ASME*, vol. 140, no. 6, 2018.

[49] P. F. G. Banfill, E. M. Szadurski, and A. M. Forster, "Deterioration of natural hydraulic lime mortars, ii: Effects of chemically accelerated leaching on physical and mechanical properties of carbonated materials," *Construction and Building Materials*, vol. 111, pp. 182–190, 2016.

[50] M. Malena, M. Sangirardi, and G. de Felice, "Steel reinforced grout under uniaxial load: Experimental evidences and numerical modelling," *Construction and Building Materials*, vol. 227, p. 116808, 2019.