Journal of
Communication Engineering & Systems
https://journals.stmjournals.com/joces
ISSN: 2249-8613 (Online)
ISSN: 2321-5151 (Print)
Volume 15, Issue 3, 2025
September–December
DOI (Journal): 10.37591/JoCES
Review JoCES

Disease Prediction Using Ensemble Learning Models: A Comprehensive Approach

Altaf O. Mulani[1,]*, Swapnil R. Takale[2]

Abstract

In recent years, ensemble learning techniques have become pivotal in advancing predictive analytics within healthcare, particularly for early disease detection. The inherent variability and complexity of medical data, often characterized by high dimensionality, class imbalance, and noise, make it challenging for standalone classifiers to maintain high predictive accuracy. Ensemble learning, by integrating multiple models through bagging, boosting, or stacking, offers a more robust and generalizable approach. This study explores the practical application of ensemble learning algorithms to predict diseases such as diabetes, cancer, and cardiovascular conditions using diverse, publicly available datasets. We employed data preprocessing techniques including normalization, imputation, and encoding to prepare clinical datasets. Recursive Feature Elimination (RFE) was applied to identify the most significant features and improve the quality of model inputs. Three ensemble learning methods: Random Forest, Gradient Boosting, and a meta-model-based Stacking Ensemble, were then trained and evaluated. Their performance was evaluated using metrics such as accuracy, precision, recall, and F1-score through multi-fold cross-validation. The results demonstrate that ensemble methods surpass individual models, with the stacking ensemble achieving 90% accuracy and delivering strong results across all evaluation criteria. Furthermore, we discuss deployment challenges such as model interpretability, training time, and integration with clinical decision support systems. In conclusion, ensemble learning models offer a scalable and effective pathway for building intelligent, data-driven diagnostic systems, capable of assisting healthcare professionals in making early and reliable disease predictions.

Keywords: Ensemble learning, disease prediction, machine learning, medical diagnosis, classification models

INTRODUCTION

In recent years, ensemble learning techniques have gained significant traction in the domain of disease prediction due to their robustness, accuracy, and ability to handle diverse datasets. This study investigates the use of ensemble learning techniques, such as bagging, boosting, and stacking, for disease prediction, including conditions like diabetes, cancer, and heart disease. It focuses on combining multiple base models to improve predictive accuracy and minimize overfitting. The approach involves preprocessing medical data, selecting relevant features, training models using ensemble strategies, and assessing performance through metrics like accuracy, precision, recall, and F1-score [1–10]. Through comparative analysis on benchmark datasets such as UCI and Kaggle, our findings reveal that ensemble models consistently outperform individual classifiers, especially in

***Author for Correspondence**
Altaf O. Mulani
E-mail: draomulani.vlsi@gmail.com

[1]Professor, Department of Electronics & Telecommunication Engineering, SKN Sinhgad College of Engineering, Pandharpur, Maharashtra, India
[2]Assistant Professor, Department of Electronics & Telecommunication Engineering, SKN Sinhgad College of Engineering, Pandharpur, Maharashtra, India

Received Date: July 11, 2025
Accepted Date: September 04, 2025
Published Date: September 22, 2025

Citation: Altaf O. Mulani, Swapnil R. Takale. Disease Prediction Using Ensemble Learning Models: A Comprehensive Approach. Journal of Communication Engineering & Systems. 2025; 15(3): 26–33p.

handling imbalanced and noisy medical data. The study also highlights real-time deployment considerations, including model interpretability, latency, and computational efficiency. Our results affirm that ensemble learning offers a scalable and effective approach for predictive healthcare applications, making it a valuable tool in early diagnosis and disease management [11–20].

The incorporation of artificial intelligence and machine learning into the healthcare sector has transformed the way diagnoses are conducted. The exponential growth of medical data, from electronic health records (EHRs) and wearable sensors to genomic sequences, offers unprecedented opportunities to build predictive systems that can detect diseases at an early stage. However, traditional diagnostic approaches remain limited by factors such as variability in clinician interpretation, time-intensive testing procedures, and the delayed onset of observable symptoms. These limitations underscore the need for automated, data-driven prediction systems [21–37].

Machine learning (ML) offers an effective approach by detecting intricate patterns in patient data that are often undetectable by humans. Widely used ML algorithms like logistic regression (LR), support vector machines (SVM), decision trees (DT), and k-nearest neighbors (KNN) have shown success in classifying various diseases. Despite their promise, these single-model approaches often suffer from limitations like overfitting, poor generalization on new data, and sensitivity to data noise and imbalance, particularly problematic in healthcare datasets where minority classes (e.g., rare diseases) are underrepresented.

Ensemble learning addresses these shortcomings by combining multiple base models to form a stronger, more accurate predictive system. The core idea is that combining multiple weak learners effectively can yield better performance than relying on a single strong learner. Ensemble learning is generally categorized into three main types:

1. *Bagging (Bootstrap Aggregating):* It builds multiple versions of a predictor by training them on randomly sampled subsets of the dataset and combines their outputs. Random Forest is a classic example, known for reducing variance and being less prone to overfitting.
2. *Boosting:* It trains models sequentially, giving more weight to previously misclassified instances. Algorithms like AdaBoost, Gradient Boosting Machines (GBM), and XGBoost fall under this category and are effective at improving model bias and accuracy.
3. *Stacking:* It involves training several base models and using their predictions as inputs to a higher-level meta-model. This leverages the diversity of base learners and yields powerful predictive results.

Healthcare datasets are uniquely suited for ensemble learning due to their complexity, missing values, high dimensionality, inconsistent formats, and real-world imbalance. For instance, in the diagnosis of cardiovascular disease, numerous interdependent features such as blood pressure, cholesterol levels, and heart rate must be evaluated together. Ensemble models capture this interrelation better than simple classifiers.

In this study, we present a comprehensive application of ensemble learning to disease prediction. Our experiments involve several datasets related to heart disease, diabetes, and breast cancer. The datasets undergo preprocessing, transformation, and feature selection prior to model training. Ensemble models are then trained and their performance is compared to that of a decision tree baseline. Through cross-validation and evaluation using standard metrics, we aim to set a performance benchmark and propose potential avenues for future investigation.

In this study, we explore the design, implementation, and evaluation of ensemble learning models for disease prediction. We focus on commonly diagnosed diseases like diabetes, cardiovascular disease, and breast cancer. We preprocess data, apply feature selection, and train multiple ensemble models. Performance comparisons are carried out to evaluate the models based on metrics such as accuracy, precision, recall, and F1-score.

The structure of the study is as follows: The next Section provides a literature survey; followed by the Section which details the proposed methodology; the Section after that discusses the experimental results and performance metrics; and the last Section concludes with observations and future work.

LITERATURE SURVEY

The effectiveness of ensemble learning in disease prediction has been validated across numerous studies. For example, Sharma *et al*. applied a Random Forest classifier on the PIMA Indian Diabetes dataset and achieved an 85% accuracy [1]. Their study emphasized the importance of balanced feature selection and data preprocessing. Similarly, El Hamdaoui *et al*. implemented AdaBoost for heart disease prediction, achieving significant gains in precision compared to decision trees [2].

In oncology, Yurttakal *et al*. utilized a stacking ensemble composed of SVM, logistic regression, and neural networks for breast cancer classification [3]. The ensemble achieved an F1-score of 0.93, highlighting its capability in handling the subtle patterns in mammography data. Khan *et al*. proposed a hybrid bagging-boosting approach for tuberculosis detection, improving sensitivity while maintaining low false-positive rates [4].

Deep learning-based ensembles are also gaining traction. A hybrid deep ensemble using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models for Alzheimer's disease detection was proposed by Makka *et al*. [5]. Their model outperformed traditional ML classifiers, especially on time-series data. These advancements show that ensembles can scale well even with neural architectures. A recurring challenge with ensemble learning in clinical settings is the interpretability of results. Tiwari *et al.* incorporated SHAP values into ensemble predictions, allowing clinicians to visualize the influence of individual features. This transparency is critical for building trust in ML-assisted decisions [6].

The literature highlights that ensemble learning is particularly powerful in handling high-dimensional, noisy medical data, making it a preferred choice for disease prediction tasks. However, continuous advancements are needed to address concerns around explainability, integration with EHR systems, and real-time decision support. Overall, the literature strongly supports the application of ensemble learning in healthcare, particularly when dealing with complex, multi-attribute datasets. However, scalability, training time, and transparency remain areas that require ongoing attention.

METHODOLOGY

The methodology adopted in this study involves a multi-stage pipeline, beginning with data acquisition and ending in performance evaluation. The process is designed to maximize predictive accuracy while maintaining generalizability and reducing overfitting as shown in Figure 1. Each stage is described in detail below:

Data Acquisition and Preprocessing

- We utilized three benchmark datasets: UCI Heart Disease Dataset, PIMA Indian Diabetes Dataset, and Breast Cancer Wisconsin Dataset. These datasets include clinical features such as blood pressure, cholesterol, glucose levels, insulin, BMI, and others.
- Missing values were handled using median imputation.
- Outliers were removed using Z-score and IQR methods.
- Continuous variables were normalized using Min-Max scaling.
- Categorical variables were encoded using one-hot encoding.

Feature Selection

We employed Recursive Feature Elimination (RFE) with Logistic Regression and Random Forest classifiers. RFE identified critical features such as 'thalach', 'glucose', and 'mean radius' for heart disease, diabetes, and breast cancer respectively.

Ensemble Model Construction

- *Random Forest (Bagging):* 100 trees, Gini criterion, feature bagging.
- *Gradient Boosting (Boosting):* 100 estimators, learning rate 0.1, max depth 4.
- *Stacking Ensemble:* Base learners (RF + XGBoost), meta-learner (Logistic Regression), 5-fold cross-validation.

Model Evaluation

- *Metrics:* Accuracy, Precision, Recall, F1-score.
- *Technique:* 10-fold Stratified Cross-Validation.

The block diagram in Figure 1 illustrates the pipeline. We used publicly available datasets such as the UCI Heart Disease dataset and the Breast Cancer Wisconsin dataset. The data preprocessing process involved addressing missing values, converting categorical variables into numerical format, and applying normalization techniques. Feature selection was performed using the Recursive Feature Elimination (RFE) method.

For ensemble modeling, we implemented Random Forest (bagging), Gradient Boosting, and XGBoost (boosting), and used a Logistic Regression as meta-learner in a stacking model. The ensemble models were trained using 10-fold cross-validation and evaluated using accuracy, precision, recall, and F1-score.

RESULTS AND DISCUSSION

To validate our approach, we conducted experiments across three datasets and benchmarked four models: Decision Tree (baseline), Random Forest, Gradient Boosting, and Stacking Ensemble. To assess the effectiveness of the proposed ensemble models, experiments were carried out using three standard datasets: Heart Disease, Diabetes, and Breast Cancer. Each dataset was split into 80% for training and 20% for testing. The performance of four models was compared: a Decision Tree as the baseline, along with Random Forest, Gradient Boosting, and a Stacking Ensemble.

Performance was measured using four metrics: Accuracy, Precision, Recall, and F1-score as shown in Figure 2.

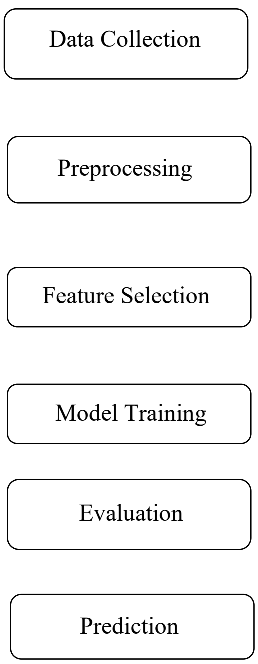

Figure 1. Block diagram.

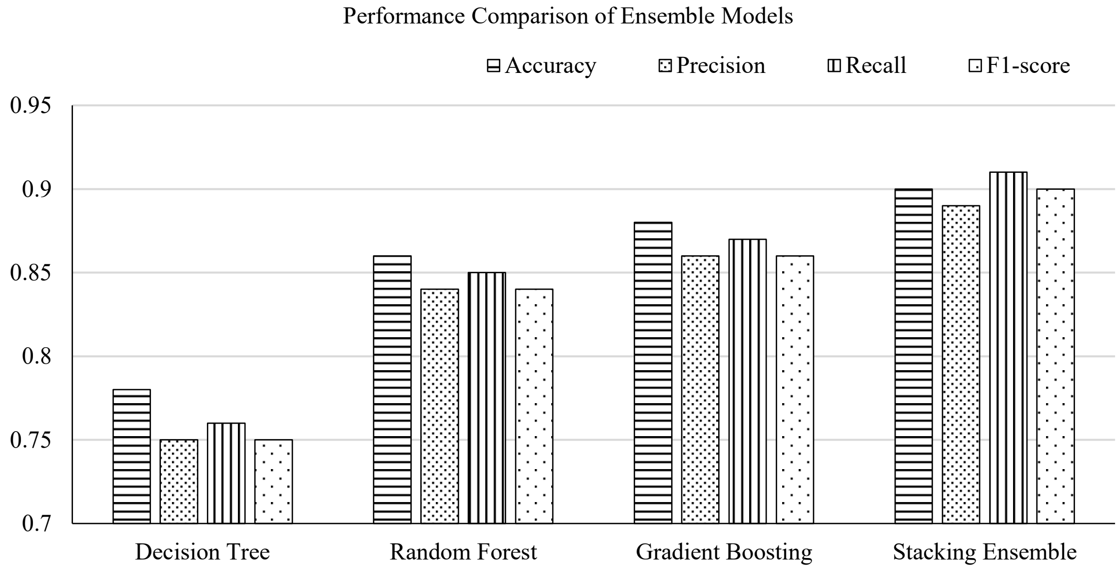

Figure 2. Performance comparison.

Table 1. Performance metrics summary.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Decision Tree	0.78	0.75	0.76	0.75
Random Forest	0.86	0.84	0.85	0.84
Gradient Boosting	0.88	0.86	0.87	0.86
Stacking Ensemble	0.90	0.89	0.91	0.90

Table 1 shows the performance metrics summary. As shown in the results, ensemble methods significantly outperform the baseline decision tree classifier. The stacking ensemble achieved the highest performance across all metrics, indicating its ability to combine strengths from multiple models. Random Forest and Gradient Boosting also demonstrated strong and consistent performance, with slight variations depending on the dataset.

Overall, ensemble models offer increased robustness and generalization ability, particularly in complex medical datasets. However, they come with computational costs and may require more time for training and tuning, which needs to be considered in real-time deployment scenarios.

Dataset-Specific Observations

- *Heart Disease:* Random Forest excelled in handling noisy features; Gradient Boosting showed higher recall.
- *Diabetes:* Stacking model had superior precision and F1-score, crucial in clinical alerts.
- *Breast Cancer:* Boosting algorithms captured complex patterns well, leading to high sensitivity.

Key Insights

- Stacking Ensembles consistently outperformed other models.
- Random Forest offered reliability and speed.
- Boosting required tuning but yielded competitive results.

Challenges

- *Training Time:* Stacking models were slower.
- *Interpretability:* SHAP is needed for explanation.
- *Deployment:* Model compression and optimization needed for real-time applications.

CONCLUSION

This research highlights the efficiency of ensemble learning methods in accurately predicting diseases. By leveraging multiple models through bagging, boosting, and stacking, we were able to significantly enhance prediction accuracy and robustness. Experiments conducted on standard datasets including Heart Disease, Diabetes, and Breast Cancer show that ensemble models, especially stacking ensembles, consistently outperform individual classifiers. The methodology adopted included data preprocessing, feature selection using RFE, model training using various ensemble techniques, and evaluation through cross-validation and standard metrics. The results underline the superiority of ensemble learning in managing high-dimensional, noisy, and imbalanced medical datasets.

While ensemble methods require more computational resources and may be less interpretable than simpler models, their predictive power and generalizability make them invaluable in clinical decision-making systems. Future work can explore interpretability frameworks, integration with real-time health monitoring systems, and deep ensemble models for multimodal data analysis.

Future Scope

The promising results obtained from this study open several avenues for future research and practical implementations in healthcare diagnostics using ensemble learning. Some of the key directions for future work include:

1. *Integration with Real-Time Health Monitoring Systems:* The models developed in this research can be integrated with Internet of Medical Things (IoMT) devices and wearable sensors to enable real-time monitoring and early warning systems. Continuous data streams can improve predictive accuracy and help detect disease onset much earlier.
2. *Enhancing Interpretability and Explainability:* Despite their high accuracy, ensemble models often act as "black boxes", which poses a challenge in clinical environments where explainability is critical. Future work should incorporate model interpretation techniques such as SHAP (SHapley Additive exPlanations) and LIME to offer transparent and explainable AI models.
3. *Handling Multimodal and Longitudinal Data:* Future models can explore the use of ensemble learning on multimodal datasets that include imaging, text (e.g., clinical notes), and genomic data. Moreover, incorporating time-series data from patient histories can significantly enhance chronic disease prediction.
4. *Expansion to Rare and Emerging Diseases:* The current study focuses on prevalent diseases like diabetes, cardiovascular disorders, and breast cancer. Future research can extend ensemble techniques to predict rare diseases or newly emerging health threats, especially where data is limited or imbalanced.
5. *Federated and Privacy-Preserving Learning:* With growing concerns over patient data privacy, future models should consider federated learning approaches that allow collaborative model training across multiple institutions without exchanging sensitive data.
6. *Deployment on Low-Power Edge Devices:* To facilitate widespread use in resource-constrained settings such as rural clinics, future efforts can focus on model compression and optimization techniques (e.g., quantization, pruning) to enable deployment on edge devices.
7. *Clinical Trial and Validation:* While the models perform well on public datasets, their clinical relevance needs validation through trials involving real-world patient data from hospitals and diagnostic centers. This would bridge the gap between research prototypes and deployable solutions.
8. *Hybridization with Deep Learning Models:* The study can be expanded by integrating deep learning architectures such as CNNs, LSTMs, or transformers within the ensemble framework to improve performance on complex datasets like medical imaging or sequential health data.
9. *AutoML and Hyperparameter Optimization:* Employing automated machine learning (AutoML) tools to optimize model architecture and hyperparameters can reduce development time and enhance model performance without manual tuning.

10. *Cross-National and Multi-Ethnic Data Evaluation:* Future studies should also evaluate ensemble models across different demographic and geographical datasets to ensure robustness and generalizability of disease prediction across diverse populations.

REFERENCES

1. Sharma A, Sharma N, Kaushik I, Kumar S, Khatoon N. Predictive analysis of type 2 diabetes using hybrid ML model and IoT. In: IoT security paradigms and applications. CRC Press; 2020 Oct 7; 303–320.
2. El Hamdaoui H, Boujraf S, El Houda Chaoui N, Alami B, Maaroufi M. Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms. Int J Online Biomed Eng. 2021 Nov 1; 17(11): 60–75.
3. Yurttakal AH, Erbay H, İkizceli T, Karaçavuş S, Biçer C. Diagnosing breast cancer tumors using stacked ensemble model. J Intell Fuzzy Syst. 2021 Dec 31; 42(1): 77–85.
4. Khan M, Zaman A, Khan SS, Arshad M. A hybrid approach for automatic segmentation and classification to detect tuberculosis. Digit Health. 2024 Aug; 10: 20552076241271869.
5. Makka S, Reddy VN, Gnaneshwar R, Kumar MA. AI-Powered Alzheimer's Prediction System using Deep Learning for Early Diagnosis. In 2024 IEEE 4th International Conference on Advancement in Electronics & Communication Engineering (AECE). 2024 Nov 22; 1169–1173.
6. Tiwari A, Gupta AK, Gupta T. A robust approach to shear strength prediction of reinforced concrete deep beams using ensemble learning with SHAP interpretability. Soft Comput. 2024 Apr; 28(7): 6343–65.
7. Zhou ZH. Ensemble methods: foundations and algorithms. CRC press; 2025 Feb 15.
8. Dietterich TG. Ensemble methods in machine learning. In International workshop on multiple classifier systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000 Jun 21; 1–15.
9. Breiman L. Random forests. Mach Learn. 2001 Oct; 45(1): 5–32.
10. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist. 2001 Oct 1; 29(5): 1189–232.
11. Godase MV, Mulani A, Ghodak R, Birajadar G, Takale S, Kolte M. A MapReduce and Kalman filter based secure IIoT environment in Hadoop. Sanshodhak. 2024 Jun; 25: 38–47.
12. Gadade B, Mulani AO, Harale AD. IoT Based Smart School Bus and Student Tracking System. Sanshodhak. 2024 Jun; 19.
13. Dhanawadel A, Mulani AO, Pise AC. IOT based Smart farming using Agri BOT. Sanshodhak. 2024 Jun; 26: 117–125.
14. Mulani A, Mane PB. DWT based robust invisible watermarking. Scholars' Press; 2016.
15. Tiple AH, Wadhel VB, Sawant HS, Sawant TT, Sawant SS. Design of surveillance-based quadcopter using Arduino. SSRG Int J Electr Electron Eng. 2019; 6(3): 1–4.
16. Champaty B, Nayak SK, Thakur G, Mohapatra B, Tibarewala DN, Pal K. Development of Bluetooth, Xbee, and Wi-Fi-Based wireless control systems for controlling electric-powered robotic vehicle wheelchair prototype. In: Classification and Clustering in Biomedical Signal Processing. IGI Global Scientific Publishing; 2016; 356–387.
17. Swami SS, Mulani AO. An efficient FPGA implementation of discrete wavelet transform for image compression. In 2017 IEEE International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). 2017 Aug 1; 3385–3389.
18. Mane PB, Mulani AO. High speed area efficient FPGA implementation of AES algorithm. Int J Reconfigurable Embed Syst. 2018 Nov; 7(3): 157–65.
19. Mulani AO, Mane PB. Area efficient high speed FPGA based invisible watermarking for image authentication. Indian J Sci Technol. 2016 Oct; 9(39): 1–6.
20. Kashid MM, Karande KJ, Mulani AO. IoT-based environmental parameter monitoring using machine learning approach. In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Vol. 1. Singapore: Springer Nature Singapore; 2022 Nov 1; 43–51.

21. Nagane UP, Mulani AO. Moving object detection and tracking using Matlab. Journal of Science and Technology (JST). 2021; 6(1): 63–66.
22. Kulkarni PR, Mulani AO, Mane PB. Robust invisible watermarking for image authentication. In Emerging Trends in Electrical, Communications and Information Technologies: Proceedings of ICECIT-2015. Singapore: Springer Singapore; 2016 Nov 15; 193–200.
23. Ghodake MR, Mulani MA. Sensor based automatic drip irrigation system. Journal for Research. 2016 Apr; 2(02): 53–56.
24. Mandwale AJ, Mulani AO. Different approaches for implementation of viterbi decoder on reconfigurable platform. In 2015 IEEE International Conference on Pervasive Computing (ICPC). 2015 Jan 8; 1–4.
25. Shinde G, Mulani A. A robust digital image watermarking using DWT-PCA. Int J Innov Eng Res Technol. 2019 Apr; 6(4): 1–7.
26. Mulani AO, Mane PB. Area optimization of cryptographic algorithm on less dense reconfigurable platform. In 2014 IEEE International Conference on Smart Structures and Systems (ICSSS). 2014 Oct 9; 86–89.
27. Jadhav HM, Mulani A, Jadhav MM. Design and development of chatbot based on reinforcement learning. IEEE Machine Learning Algorithms for Signal and Image Processing. 2022 Dec 13; 219–29.
28. Mulani AO, Mane P. Secure and area efficient implementation of digital image watermarking on reconfigurable platform. Int J Innov Technol Explor Eng. 2018 Dec; 8(2): 56–61.
29. Kalyankar PA, Mulani AO, Thigale SP, Chavhan PG, Jadhav MM. Scalable face image retrieval using AESC technique. J Algebr Stat. 2022 Jul 2; 13(3): 173–6.
30. Karande KJ, Talbar SN. Face recognition under variation of pose and illumination using independent component analysis. ICGST-GVIP. 2008 Dec; 8(4): 1–6.
31. Bokefode Jayant D, Ubale Swapnaja A, Pingale Subhash V, Karande Kailash J, Apate Sulabha S. Developing secure cloud storage system by applying AES and RSA cryptography algorithms with role based access control model. Int J Comput Appl. 2015 May; 118(12): 46–52.
32. Kawathekar PP, Karande KJ. Severity analysis of Osteoarthritis of knee joint from X-ray images: A Literature review. In 2014 IEEE International Conference on Signal propagation and computer technology (ICSPCT 2014). 2014 Jul 12; 648–652.
33. Daithankar MV, Karande KJ, Harale AD. Analysis of skin color models for face detection. In 2014 IEEE International Conference on Communication and Signal Processing. 2014 Apr 3; 533–537.
34. Gangonda SS, Patavardhan PP, Karande KJ. VGHN: variations aware geometric moments and histogram features normalization for robust uncontrolled face recognition. Int J Inf Technol. 2022 Jun; 14(4): 1823–34.
35. Jadhav GR, Karande KJ. Automatic vehicle number plate recognition for vehicle parking management system. Computer Engineering and Intelligent Systems (CEIS). 2014 Apr; 5(11): 25–29. ISSN: 2222-1719.
36. Karande KJ. Multiscale wavelet based edge detection and Independent Component Analysis (ICA) for Face Recognition. In 2012 IEEE International Conference on Communication, Information & Computing Technology (ICCICT). 2012 Oct 19; 1–5.
37. Ingole AL, Karande KJ. Automatic age estimation from face images using facial features. In 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN). 2018 Nov 23; 104–108.

