International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

A stacked ensemble machine learning framework for healthcare risk prediction

Sudharson D [1], Sri Dharaneesh P [1], Preethi K N [1], Rakshitha S [1 *], Vijilesh V [2] and Vimal EA [3]

[1] *Department of Artificial Intelligence and Data Science, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India.*
[2] *Department of MCA, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India.*
[3] *Department of CSE, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India.*

International Journal of Science and Research Archive, 2026, 19(01), 827-834

Publication history: Received on 02 March 2026; revised on 18 April 2026; accepted on 20 April 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.1.0824

Abstract

Accurate prediction of patient health risk is important for early diagnosis and clinical decision-making. However, healthcare data is often complex and noisy, which limits the performance of individual machine learning models. This study proposes a stacked ensemble framework for healthcare risk prediction. The approach combines Random Forest, LightGBM, XGBoost, and CatBoost as base learners, with Logistic Regression as a meta-learner. The model is trained on 49,942 patient records with 15 clinical and lifestyle features. Class imbalance is handled using SMOTE applied within a stratified 5-fold cross-validation framework. SHAP-based explainability is used to improve interpretability of model predictions along with preprocessing steps such as feature scaling and encoding. Experimental results show strong performance across all evaluation metrics, outperforming individual models.

Keywords: Stacked ensemble; Healthcare risk prediction; Machine learning; SMOTE; Clinical decision support

1. Introduction

Healthcare data includes patient records, lab results, clinical measurements, and lifestyle information, and its volume is continuously increasing with the adoption of digital systems in healthcare. This data is mainly used for early disease detection, risk assessment, and clinical decision-making. However, it is often inconsistent, contains missing values, and shows strong class imbalance, making reliable prediction difficult.

Machine learning is widely used in this area because it can learn patterns from complex datasets. However, individual models often struggle with noisy and imbalanced data, and their performance can vary across datasets. While ensemble methods help improve performance, many existing approaches focus mainly on model combination rather than addressing data-level issues such as imbalance handling and consistency of evaluation.

In this study, we propose a stacked ensemble framework using Random Forest, LightGBM, XGBoost, and CatBoost as base learners, with Logistic Regression as the meta-learner. The data is pre-processed through standard cleaning steps, including handling missing values, feature scaling, and encoding. SMOTE is applied within a stratified cross-validation framework to handle class imbalance while avoiding data leakage. SHAP is used to explain feature contributions and improve interpretability.

The aim of this work is to build a model that performs well and remains reliable on real-world healthcare data, supporting clinical decision-making. Unlike many existing approaches, the framework integrates preprocessing, imbalance handling, and model training within a unified evaluation setup to ensure more consistent results.

* Corresponding author: Rakshitha S

2. Related work

Machine learning has been widely applied in healthcare for disease prediction and risk analysis. Early approaches relied on traditional classifiers such as logistic regression, decision trees, and SVMs, but these models often struggle to capture the complex, non-linear patterns found in clinical data.

Ensemble methods addressed many of these shortcomings. Random Forest and gradient boosting models like XGBoost and LightGBM have become standard choices for structured healthcare data. Studies by Ahmad et al. [1] and Gul et al. [2] confirm that ensemble approaches generally outperform individual classifiers across a range of prediction tasks.

More recent work has moved beyond accuracy toward questions of reliability and calibration. Van Calster et al. [7] and Jiang et al. [15] argue that well-calibrated probability estimates are essential in clinical settings, while Kuleshov et al. [9] and Guo et al. [10] point out that many models tend to produce overconfident predictions when outputs are used to guide patient care. Class imbalance remains a persistent challenge in healthcare datasets. SMOTE [16] is widely used for handling class imbalance. However, when applied before cross-validation, it can introduce information leakage into the validation process, leading to overly optimistic performance estimates.

Beyond this, many studies still treat preprocessing, balancing, and model training as independent stages, which reduces consistency and makes results harder to reproduce in real clinical conditions.

2.1. Research Gap

Three key limitations are observed in the existing literature. First, many studies use fragmented pipelines where preprocessing, resampling, and model training are performed separately, which can lead to inconsistencies and potential data leakage. Second, only a few works integrate heterogeneous stacked ensembles with imbalance-aware learning within a single cross-validation framework. Third, the stability of models across validation folds is often not reported, even though it is important for assessing clinical reliability.

To address these issues, our paper proposes a unified framework in which preprocessing, SMOTE, and model training are integrated within a stratified cross-validation pipeline, ensuring more consistent and reliable evaluation results.

3. Materials and Methods

Reliable healthcare prediction depends not only on which models are chosen, but also on how the data is prepared, balanced, and evaluated. In many existing approaches, these steps are handled independently, which can introduce inconsistencies and overly optimistic results. In contrast, the model integrates dataset preparation, preprocessing, imbalance handling, and model training into a single unified pipeline to ensure consistent and reproducible evaluation.

3.1. Dataset Description

The dataset used in this framework contains 49,942 patient records with 15 clinical and demographic features, including laboratory measurements such as glucose, cholesterol, and HbA1c, biometric indicators like BMI and blood pressure, and demographic attributes such as age and sex. The target variable is binary showing whether a patient is at risk (1) or not at risk (0).

The dataset exhibits class imbalance, with approximately 89.1% of samples belonging to the low-risk (negative) class and 10.9% to the high-risk (positive) class, resulting in an 8:1 distribution. The positive class represents patients at clinical risk requiring closer medical attention, while the negative class represents those with lower or no immediate risk. This setup reflects real-world screening scenarios where early detection of high-risk cases is prioritized.

Table 1 Summary of features and their clinical significance

Feature	Range	Type	Clinical Significance
Glucose (mg/dL)	52 – 208	Lab	Primary diabetes indicator; >125 suggests impaired fasting
Cholesterol (mg/dL)	90 – 352	Lab	CVD risk marker; >240 considered high
BMI	10 – 42	Biometric	Obesity indicator; >30 classified as obese

Age (years)	18 – 89	Demographic	Risk increases with age across metabolic disorders
LDL (mg/dL)	8 – 290	Lab	Atherogenic cholesterol; >160 clinically elevated
HDL (mg/dL)	10 – 90	Lab	Protective cholesterol; <40 increases CVD risk
Triglycerides (mg/dL)	30 – 400	Lab	Linked to metabolic syndrome
Systolic BP (mmHg)	90 – 180	Biometric	Hypertension criterion; >140 classified as stage 2
Diastolic BP (mmHg)	60 – 120	Biometric	Complementary to systolic in hypertension staging
Fasting Insulin (µIU/mL)	2 – 50	Lab	Insulin resistance proxy; elevated in pre-diabetes
HbA1c (%)	4.5 – 9.5	Lab	>6.5 diagnostic for diabetes
Creatinine (mg/dL)	0.4 – 2.5	Lab	Renal function marker; elevated suggests kidney stress
ALT (U/L)	10 – 95	Lab	Liver enzyme; elevated with metabolic dysfunction
Heart Rate (bpm)	50 – 120	Biometric	Resting HR links to cardiovascular fitness and stress
Sex	0 / 1	Demographic	Binary encoded: Male = 1, Female = 0

3.2. Data Preprocessing

Before training, numerical features were standardized using z-score normalization so that variables on different scales contribute equally. Categorical variables such as sex were encoded into numerical form. To ensure a fair evaluation, all preprocessing transformations were learned from the training data and then applied to validation and test sets, preventing data leakage. The dataset was split into 80% training and 20% testing using stratified sampling to preserve the original class distribution. The distribution of key features across patient groups is shown in Fig. 1.

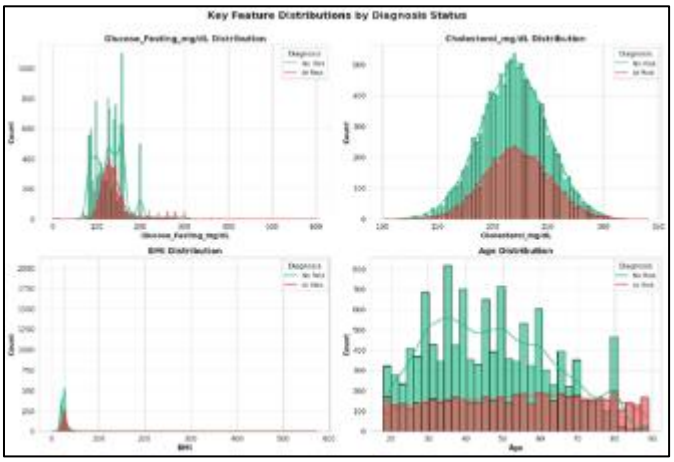

Figure 1 Distribution of key clinical features across patient groups

3.3. Handling Class Imbalance

The dataset exhibits class imbalance, where approximately 89.1% of samples belong to the low-risk (negative) class and 10.9% correspond to the high-risk (positive) class. In this study, the positive class represents patients at clinical risk, which is the minority class of primary interest. This reflects real-world screening scenarios where high-risk patients are relatively rare but critically important to detect.

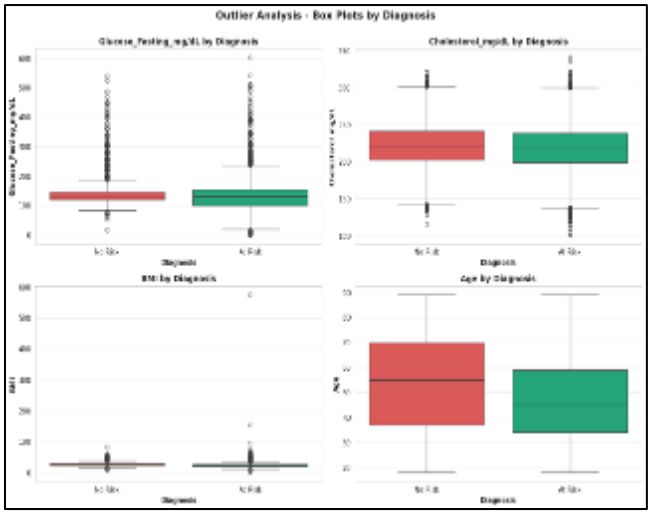

Figure 2 Outlier analysis of key features before preprocessing, indicating the presence of extreme values in the raw dataset

3.4. Model Development

The proposed framework uses a stacked ensemble consisting of four tree-based base learners - Random Forest, LightGBM, XGBoost, and CatBoost, selected for their complementary strengths in handling structured healthcare data. Each base model generates probability outputs, which are then used as inputs to a Logistic Regression meta-learner. This allows the model to learn how to weight each base learner dynamically, rather than relying on fixed or equal contributions.

Unlike traditional implementations, model training is tightly coupled with preprocessing and imbalance handling within the same pipeline, reducing variability across different data splits and improving overall robustness.

3.5. Experimental Setup and Evaluation

Training was performed using 5-fold stratified cross-validation, with SMOTE applied independently within each training fold. This ensures that oversampling does not influence validation results and avoids overly optimistic performance estimates.

The final model was evaluated on a separate held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC, providing a comprehensive assessment of performance across both majority and minority classes.

3.6. Model Interpretability using SHAP

SHAP (SHapley Additive exPlanations) is used to interpret the predictions of the proposed stacked ensemble model. After training, SHAP values are computed to quantify the contribution of each feature toward the final prediction. This allows both global interpretabilities, by identifying the most impactful clinical features across the dataset, and local interpretability, by explaining individual patient-level predictions. The use of SHAP ensures that the model is not only accurate but also transparent, which is essential for healthcare decision-support systems.

4. Results and Discussion

This section presents the performance of the proposed stacked ensemble and compares it against individual base models, with particular attention to consistency and reliability under realistic evaluation conditions.

4.1. Comparison of Individual and Stacked Models

Base models were evaluated independently before assessing the full stacked ensemble. Results are shown in Fig. 3 and Fig. 4. The quantitative performance of the top-performing models is summarized in Table 2.

Table 2 Performance comparison of top-performing models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Stacking Ensemble	0.9482	0.9558	0.8726	0.9123	0.9753
Random Forest	0.9474	0.9613	0.8643	0.9102	0.9731
LightGBM	0.9465	0.9583	0.8643	0.9089	0.9795
XGBoost	0.9429	0.9405	0.8698	0.9038	0.9742

LightGBM and XGBoost perform strongly on their own, but their results vary depending on the data split and evaluation metric. In contrast, the stacked ensemble shows more consistent performance across accuracy, recall, F1-score, and ROC-AUC. This improvement is mainly due to the ensemble combining predictions from multiple models, which reduces the effect of variability from individual learners.

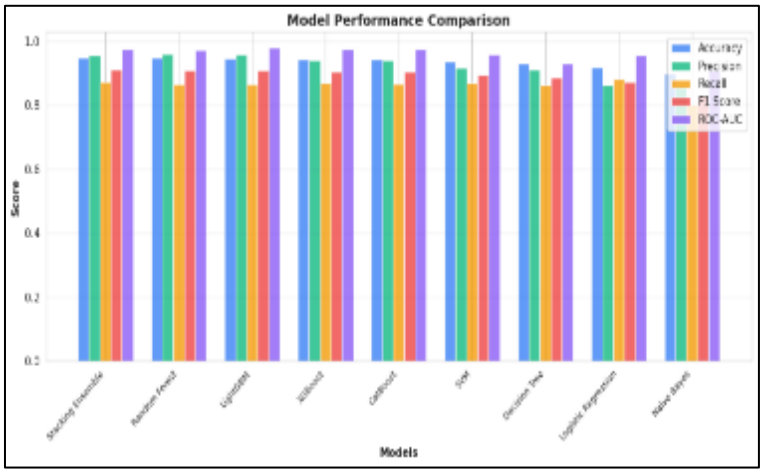

Figure 3 Performance comparison of individual machine learning models

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
8	Stacking Ensemble	0.9482	0.9558	0.8726	0.9123	0.9753
2	Random Forest	0.9474	0.9613	0.8643	0.9102	0.9731
5	LightGBM	0.9465	0.9583	0.8643	0.9089	0.9795
6	XGBoost	0.9429	0.9405	0.8698	0.9038	0.9742
7	CatBoost	0.9426	0.9414	0.8680	0.9032	0.9761
4	SVM	0.9358	0.9170	0.8708	0.8933	0.9581
1	Decision Tree	0.9318	0.9118	0.8625	0.8865	0.9318
0	Logistic Regression	0.9199	0.8614	0.8827	0.8719	0.9568
3	Naive Bayes	0.8987	0.8604	0.8020	0.8302	0.9084

Figure 4 Performance comparison between individual models and the stacked ensemble

4.2. ROC Curve Analysis

ROC curves for all models are shown in Fig. 5. The stacked ensemble achieves the highest AUC, reflecting a stronger ability to separate at-risk from non-risk patients across different thresholds. It also maintains a more stable true positive rate across different thresholds as the false positive rate increases, something individual models struggle with. In a clinical context this matters considerably, since failing to identify a high-risk patient is rarely an acceptable outcome. The results also speak to the integrity of the evaluation. Because SMOTE was applied only inside training folds, synthetic samples never influenced validation results, making the AUC figures a more honest estimate of real-world performance.

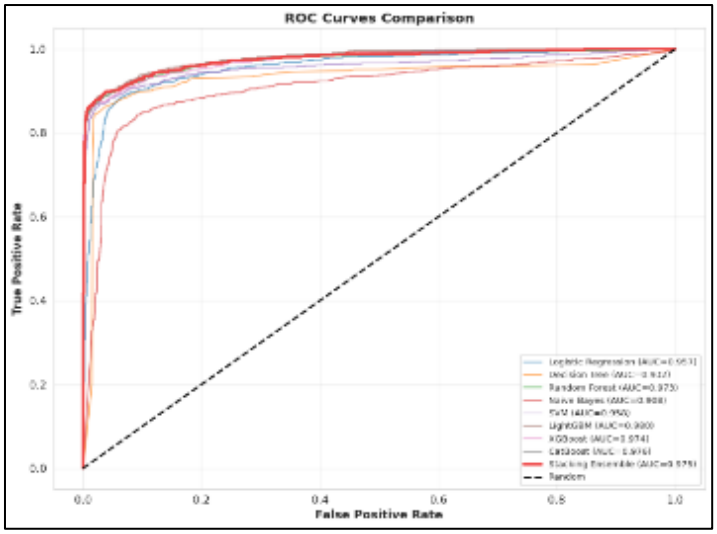

Figure 5 ROC curves showing classification performance of the models

4.3. Error Analysis using Confusion Matrix

The confusion matrix in Fig. 6 provides a closer look at where predictions succeed and fail. The most meaningful finding here is the reduction in false negatives relative to individual models. In clinical terms, a false negative means a high-risk patient is sent away as low-risk, potentially delaying diagnosis and treatment. The ensemble reduces these cases more effectively by combining predictions from multiple base models, while balanced training through SMOTE ensures that minority class patients are not frequently missed.

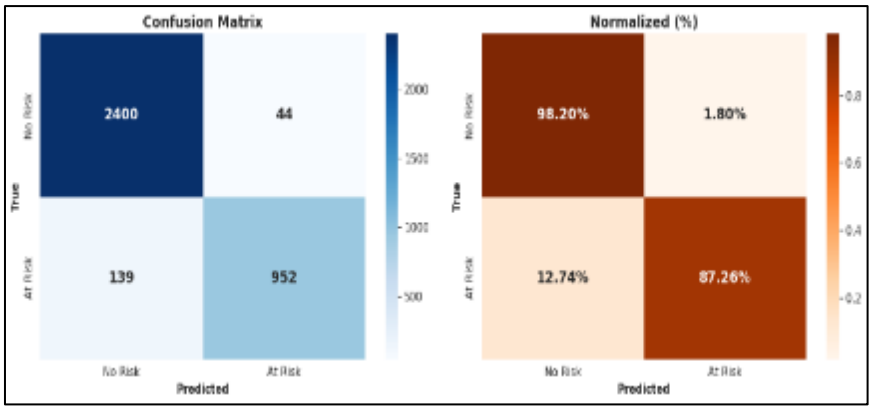

Figure 6 Confusion matrix of the proposed stacked ensemble model

4.4. Limitations and Future Work

The framework has been tested on a single structured dataset, and how it performs on data from different hospitals or clinical systems is still an open question. Interpretability is another limitation, strong predictions are only part of what clinicians need; understanding why a patient is considered as high-risk matters just as much for trust and adoption. Future work will focus on validating the model across multiple datasets, integrating SHAP explanations more deeply into the clinical workflow, and exploring whether the framework can be adapted for practical or personalised patient monitoring.

5. Conclusion

This study proposed a stacked ensemble framework for healthcare risk prediction, combining Random Forest, LightGBM, XGBoost, and CatBoost as base learners with Logistic Regression as the meta-learner. Evaluated on huge patient records, the model produced strong and consistent results across all major metrics. The key contribution of this work is the unified pipeline, integrating preprocessing, SMOTE-based balancing, and model training within stratified

cross-validation, which reduces the inconsistencies that arise when these stages are treated independently. SHAP-based explainability was also included to make model outputs more interpretable and suitable for clinical use.

A limitation of our framework is that the evaluation was conducted on a single dataset. Future work will focus on testing the framework on external clinical datasets to test applicability, as well as exploring deployment in real-time clinical decision support environments.

Compliance with ethical standards

Acknowledgements

The authors would like to thank Kumaraguru College of Technology for supporting this research work.

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

[1] Ahmad S, Hussain S, Arif M. Early-stage diabetes prediction using a stacked ensemble model. Biomed Pharmacol J. 2026;19(1).

[2] Gul G, et al. Machine learning and ensemble methods for cardiovascular disease prediction: A systematic review. Computers. 2026;15(2):45.

[3] Xu Z. A SHAP-based explainable multi-level stacking ensemble learning method for predicting the length of stay in acute stroke. arXiv preprint arXiv:2505.24101. 2025.

[4] Dharaneesh S, Preethi PKN, Rakshitha S, Sudharson D, Diwakaran M, Saravanan A. Designing intelligent and integrated analytics frameworks for personalized healthcare. In: Proc 4th Int Conf Applied Artificial Intelligence and Computing (ICAAIC). IEEE; 2025. p. 334–340.

[5] Yan AP, Guo LL, Inoue J, Patel SM, Chen YF. A roadmap to implementing machine learning in healthcare: from concept to practice. Front Digit Health. 2025.

[6] Abbas SR, Abbas Z, Zahir A, Lee SW. Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with IoT integration. Healthcare. 2024;12(24):2587.

[7] Van Calster B, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.

[8] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proc 22nd Int Conf Machine Learning (ICML). 2005. p. 625–632.

[9] Kuleshov V, Fenner N, Ermon S. Accurate uncertainties for deep learning using calibrated regression. In: Proc 35th Int Conf Machine Learning (ICML). 2018. p. 2796–2804.

[10] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proc 34th Int Conf Machine Learning (ICML). 2017. p. 1321–1330.

[11] Rajkomar R, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–1358.

[12] Ke G, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Adv Neural Inf Process Syst (NeurIPS). 2017;30:3146–3154.

[13] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc 22nd ACM SIGKDD Int Conf Knowledge Discovery and Data Mining. 2016. p. 785–794.

[14] Obermeyer Z, Emanuel EJ. Predicting the future—Big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–1219.

[15] Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. J Am Med Inform Assoc. 2012;19(2):263–274.

[16] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357.

[17] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proc 8th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining. 2002. p. 694–699.

[18] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Adv Large Margin Classifiers. MIT Press; 1999. p. 61–74.

[19] Dorai D, et al. A novel machine learning approach for software reliability growth modelling with pareto distribution function. Soft Computing. 2019;23:8379–8387.

[20] Vignesh K, et al. Classification of diabetics and cardiovascular diseases using machine learning frameworks. International Journal of Research Publication and Reviews. 2022;3:1848–1853.

[21] D, et al. Enhancing the efficiency of lung disease prediction using CatBoost and expectation maximization algorithms. In: Proc 4th Int Conf Inventive Research in Computing Applications (ICIRCA). 2022. p. 57–61.

[22] Dr, et al. Performance analysis of enhanced AdaBoost framework in multifacet medical dataset. Natural olatiles & Essential Oils. 2021;8:1752–1756.

[23] S. Ashfia Fathima, et al. Performance evaluation of improved AdaBoost framework in randomized phases through stumps. IEEE Xplore. 2021. p. 1–6.

[24] D, et al. Self-reliant dimensionality reduction using improved Pareto distribution PCA framework. IEEE Xplore. 2021. p. 1–5.

[25] D, et al. A novel AI framework for assuring data sustainability in healthcare dataset. In: Proc 3rd Int Conf Ubiquitous Computing and Intelligent Information Systems (ICUIS). 2023. p. 161–166.

[26] D, et al. Cloud-enabled predictive modeling of cancer progression in digital twins: A LightGBM classification approach. Smart Data Intelligence. 2024.

[27] D, et al. AI powered monitoring and risk prediction for maternal health to ensure fetal well-being. In: Proc ICAECA. 2025. p. 1–5.

[28] Kanagaraj, et al. Deep learning techniques to analyze chest X-ray and provide accurate predictions regarding lung disease. In: Proc ICAECA. 2025. p. 1–6.

[29] Vignesh et al., “Impact of Classification Algorithms on Cardiotocography Dataset for Fetal State Prediction,” *Asian Journal of Computer Science Engineering*, vol. 7.0, pp. 71-76, 2023.

[30] D et al., “Implementing Catboost Algorithm for Allergen Cross Contamination Detection in Food Industry,” *2023 International Conference on Emerging Research in Computational Science (ICERCS)*, vol. nan, pp. 1-4, 2024.