A Statistical Analysis of Outliers in Diabetes Prediction

Gurleen Kaur[1] Harsh Pratap Jain[2] Vashnavi Walia[3] Asmita Singh[4]
[1,2,3,4]Department of Computer Science Engineering
[1,2,3,4]AIT, Chandigarh University, Mohali, India

Abstract — In response to the growing incidence of the disease, this study creates a machine learning pipeline for diabetes pre diction utilizing the Pima Indians Diabetes database. This paper presents the analysis of outliers and its effects on the performance of prediction model. In this study, we included Naive Bayes, Random Forest, Decision Trees, KNN, and Logistic Regression. The analysis gives the selection of optimal predictive method, significant features identification, and the effects of class imbalance. Based on clinical factors such as blood pressure, insulin, glucose, and BMI, we performed this study. Preprocessing the data is a part of the pipeline to handle outliers and missing values and provide objective training. Performance measurements are used, with an emphasis on F1-score because of class imbalance. These metrics include accuracy, precision, recall, and F1-score. We analyze the results on the basis of these performance parameters and concluded that Random Forest attaining the maximum accuracy and F1 score.

Keywords: Diabetes Prediction Model, Outliers, Machine Learning, Naive Bayes, Random Forest, Decision Trees, KNN, Logistic Regression, F1-score, Confusion Matrix

I. INTRODUCTION

The International Diabetes Federation (IDF) estimates that 537 million persons *(H.Sun et. Al., 2022)* worldwide will have diabetes in 2021, making it one of the most common chronic diseases. Diabetes, which is characterized by raised blood sugar levels (hyperglycemia), is caused by the body either becoming resistant to insulin (Type 2 diabetes) or not producing enough insulin (Type 1 diabetes) *(C. G. Awuchi et. Al., 2020).* Unmanaged diabetes has serious long- term repercussions that can cause complications like renal failure, nerve damage, cardiovascular disease, and in some cases, death. As a result, avoiding these negative effects requires early diabetes testing and treatment. This has prompted scientists and medical experts to investigate cutting edge technologies, in particular machine learning (ML), to enhance diabetes diagnosis and prediction.

The healthcare sector underwent a complete transformation because machine learning enables extensive data analysis that produces exact predictions. Predictive modelling learning delivers improved characteristics of speed and accuracy and enhanced scalability functions over traditional statistical tools. Our study investigates this direction by developing a machine learning pipeline which predicts diabetes status through clinical data from Pima Indian Diabetes patients.

Our study will especially focus on six famous machine learning algorithms: AdaBoost, Naive Bayes, K-Nearest Neighbours (KNN), Logistic Regression, and Decision Trees *(K. M. Rani, 2020).* Two sophisticated algorithms Random Forest and AdaBoost achieve better accuracy by applying ensemble decision trees or boosting mechanics while the simpler Naive Bayes algorithm delivers interpretability with its approach. The KNN algorithms utilizing instance –based learning approach have been paired with logistic regression as a statistical procedure for binary classification to address diverse data distribution requirements. The interpretability of decision trees transforms them into crucial clinical decision – making instruments.

We seek to identify the most suitable algorithm from the available dataset that will work best for diabetes prediction while prioritizing vital healthcare performance metrics comprising accuracy, recall and precision. This work evaluates the suitability of predictive diagnostic methods for future applications while highlighting their benefits and drawbacks.

In conclusion, the presented study seeks to contribute knowledge that links machine learning with healthcare applications. We aim to show how predictive modeling works for diabetes diagnosis through building a robust system that first prepares clinical data while training various machine learning algorithms to execute comprehensive performance evaluations of each model system. The research findings present opportunities for medical practice improvement through accurate data-driven tools that identify diabetes risks before patients need early medical intervention to enhance treatment outcomes.

II. LITERATURE REVIEW

A. Machine Learning Techniques for Predicting Diabetes

It is commonly known that diabetes prediction is done using various machine learning algorithms. Mishra's study focused, in diabetes prediction, on the use of data-driven techniques which utilizes medical datasets (*Mishra, 2024*), in order to compare algorithms like Decision Trees, K-Nearest Neighbors, and SVMs in terms of accuracy and precision. By concluding that early diagnosis through these algorithms, the study shows the significance of data-driven approaches for health diagnostics allowing better illness management and intervention tactics (*S.Gamage, 2024*).

To facilitate early diabetes detection, an EHR system applying ML algorithms was described as one of *Jagadamba et al. (2024)* developing works. This investigation implements a systematic method to enhance health data storage and analysis because these processes enable the identification of risk variables while enabling rapid intervention. Basically, the findings presented by Jagadamba et al. demonstrate that linking EHR data to current machine learning algorithms produces better early diagnoses and improved patient results.

B. Healthcare Interpretable Models

Machine learning interpretability stands as a vital healthcare element because it influences the usefulness as well as credibility of clinical models. The healthcare industry requires precise balancing between model accuracy and interpretability since it depends on understanding how the model makes decisions for safe clinical implementations. Healthcare professionals depend on interpretability to utilize

ML algorithms properly in fields that require medical decision-making throughout diagnosis and treatment planning and patient care (*Lipton, 2018*).

Interpretable accuracy in healthcare machine learning models remains essential for medicine because healthcare professionals require clear predictions to base their clinical decisions on model predictions. By stressing the importance of medical practitioners needing algorithms, McKelvey performed an in-depth study of interpretable machine learning models which they can trust in crucial decision-making tasks (*McKelvey, 2018*).

C. A Comparative Study of Machine Learning Algorithms

For diabetes complications prediction the research by Abaker about machine learning algorithms reveals how SVM, Random Forest and Decision Trees demonstrate both strengths and limitations in diabetes management applications (*Abaker, 2024*). Clinical decision-making benefits from clever model selection as well as from models that deliver explanations of their analysis. Using diabetic data, several algorithms receive evaluation through their performance assessment which exhibits different levels of success rates and accuracy.

D. Evaluation Metrics

Accuracy and precision with recall together forming the F1 score exists among the evaluation measures for machine learning models. The role of a vital evaluation metric which unites recall and precision to produce balanced model performance assessments is served by F1 score, mostly during non-uniform class distribution as explained in *Manning et al. (2008)*. The correct prediction of diabetes calls this parameter to ensure proper predictions availability because class imbalance affects both model application and reliability.

The article explains accuracy and precision recall, and F1 score significantly to show their individual influences on model assessment (*Bird, 2013*). A harmonious evaluation point between precision and recall is created by F1 score, while recall measures authentic positive detection validity and precision illustrates correct positive predictions. The evaluation measures provide comprehensive performance recognition that makes them specifically beneficial when operating with unbalanced datasets and when detecting conditions like diabetes.

In conclusion, research indicates fundamental developments regarding machine learning-based diabetes prediction through improving early detection and healthcare results according to the literature. The evaluation metrics F1-score together with precision, recall and accuracy are vital for healthcare applications because they help achieve better model performance and deliver precise and balanced predictions especially when interpretability matters. The potency of Decision Trees, SVM and Neural Networks in diabetes prediction is shown in research but investigation now prioritizes on making these models more understandable and scalable.

III. Methodology

A. Data Collection:

Dataset was taken from Kaggle which is publically available on Kaggle dataset repository. Thorough list of medical records with elements including features like:

- Glucose levels: A critical indicator for diabetes diagnosis is also included in this collection for research.
- Insulin levels: Essential to understanding how the sugar is handled by our body.
- Blood Pressure: Diabetes and high blood pressure are constantly affiliated with each other.
- Body Mass Index (BMI): A significant critical characteristic for diabetes is obesity, indicated on the basis of one's Body Mass Index (BMI).
- Other Features: On availability, variables such as family, gender, and age history of diabetes were likewise gathered.

The outcome label to justify whether the patient was diabetic or not, was binary, i.e. 1 indicating for diabetic and 0 indicating for non-diabetic, which was overcome as the goal variable of the supervised machine learning model. For excluding data quality, integrity and corrupted or incomplete records, a verification method was guaranteed

B. Data Preprocessing:

To ensure for model construction, that the data is ready, preprocessing is a key phase. The following tasks were completed:

- Missing Value Imputation: A few records consisted of some missing values, particularly for crucial variables like "BMI" and "Insulin". Since missing values can skew model training or result in errors, thus were treated with full care. The mean values for "insulin" and "BMI", categorized by the diabetes result (diabetic vs. non-diabetic), were used to ascribe missing values. The ascribed values are guaranteed, for every class, to match the general distribution of data for every class by using this technique.
- Outlier Detection and Treatment: The features such as 'Glucose' and 'Insulin' used the Interquartile Range (IQR) approach to identify outliers. Outlier values were defined as those that fell outside of the first and third quartiles, or 1.5 times the IQR. The outliers were either capped or eliminated to avoid skewing the model, at the top and lower boundaries.
- Normalization/Standardization: To stop features with higher magnitudes (like Glucose), standardization was used, from taking the middle stage in the models like Logistic Regression and KNN. As a result, the characteristics had a zero mean and a one standard deviation. Normalization significantly increased the performance of Decision Trees and Random Forests, even though it was not necessary for gradient-based algorithms such as Logistic Regression and AdaBoost.

C. Feature Selection:

In order to prevent regression to the mean, and enhance interpretability, lower model complexity, feature selection was required since several features contributed to the prediction of diabetes. The ensuing actions were performed:

– Random Forest Classifier for Feature Importance: To rank the characteristics according to how well they contributed to prediction accuracy, a 'Random Forest' classifier was employed. To determine how important each feature is, the classifier calculates how much the prediction error rises when the data is shuffled for a given feature. The identification of critical features such as insulin, BMI, and glucose levels was done, whereas the removal or reduction of less relevant elements (such skin thickness) was studied.
– Taking an example as shown in Fig.3, this bar graph illustrates each feature's significance based on the Random Forest classifier, demonstrating how each feature affects model prediction.
– Correlation Analysis: In order to understand the correlations between features, a correlation matrix was produced. Factors with higher collinearity, i.e. the ones which were correlated highly with each other, were carefully examined, since strongly correlated features lowers the model's efficiency. One of the correlated features can be changed or eliminated in such situations.

D. Modelling:

Six machine learning algorithms were chosen, each representing a distinct classification strategy, to predict the diabetes outcome:

– Logistic Regression: To estimate the probability of a binary outcome, a linear model as a logistic function is used. It is explainable and works well when the features are linearly separable.
– K-Nearest Neighbors (KNN): It is a non- parametric, instance-based learning algorithm. On the basis of the majority class among its nearest neighbor, it classifies an individual in the feature space.
– Decision Trees: A tree-based algorithm that splits the dataset into subsets based on feature values. It is explained able and can handle non-linear relationships but is liable to over-optimization.
– Random Forest: An ensemble technique that builds several decision trees and combines the predictions of each tree. By lowering variance and overfitting, Random Forest produces forecasts that are more reliable.
– Adaptive Boosting, or AdaBoost: A boosting technique that creates a powerful classifier by combining weak classifiers, such as shallow decision trees. By modifying weights according to prior models' faults, it increases prediction accuracy in challenging situations.
– Naive Bayes: A statistical classifier based on the assumption that features are independent. This assumption is simple, but it can still work well, particularly when features are largely independent.

E. Model Evaluation:

In order to determine the quality of the categorization, performance metrics were used to analyze the models. The subsequent actions were executed:

– Train-Test Split: A 20% test set and an 80% training set were created from the dataset. By dividing the data, this split made sure the models were trained on most of the data but assessed on data that hadn't been seen, giving an objective assessment of performance.
– Cross-Validation: To minimize overfitting and assure generalizability, k-fold cross-validation (with k=5 or 10) was applied. This separates the training data into k subgroups, cycles through all subsets, trains the model on k-1 subsets, then tests it on the remaining subset.
– Evaluation Metrics:
  • F1 Score: An equilibrium between recall and precision, particularly helpful in unbalanced datasets. It evaluates how well the test detects diabetic people (1) without placing undue emphasis on any particular metric.
  • Accuracy: The proportion of instances that were accurately predicted to all instances. Even though accuracy is a widely used metric, unbalanced data may cause it to fail to accurately reflect the model's performance.
  • Confusion Matrix: To evaluate the performance of the model across several classes, a matrix was created to show true positives, true negatives, false positives, and false negatives.

1) Logistic Regression
2) K-Nearest Neighbours
3) Decision Tree
4) Random Forest
5) AdaBoost
6) Naive Bayes

IV. IMPLEMENTATION

A. Outlier Detection

Box Plots for All Features: The following box plots show the distribution of every feature in the data. These box plots assist in the identification of the spread and possible outliers within the data for every feature and offer insights into the central tendency and variability of the values. The box plots are created by using Plotly with a high-contrast color scheme for clearer visualization.

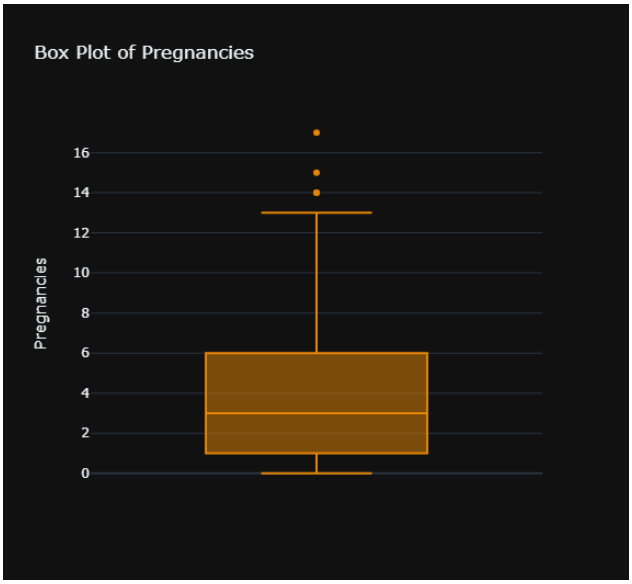

Fig. 1: Outlier Detection in Pregnancies

A box plot presents the data in 4 regions Lower Fence to Q1, Q1 to Median, Median to Q3, Q3 to Upper

Fence. Any data point which is below the lower fence will be considered as outlier. Similarly any data point which is above the upper fence will be considered as outlier. These outliers or data points will be removed from the dataset in the data preprocessing part.

B. Box Plot of Insulin

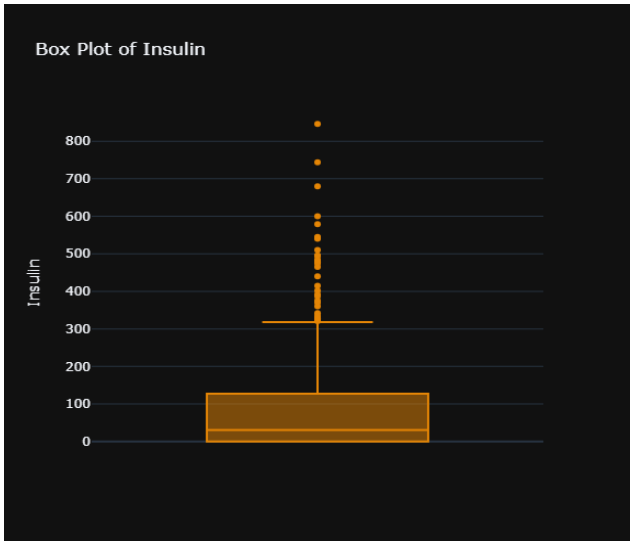

Fig. 2: Outlier Detection in Insulin

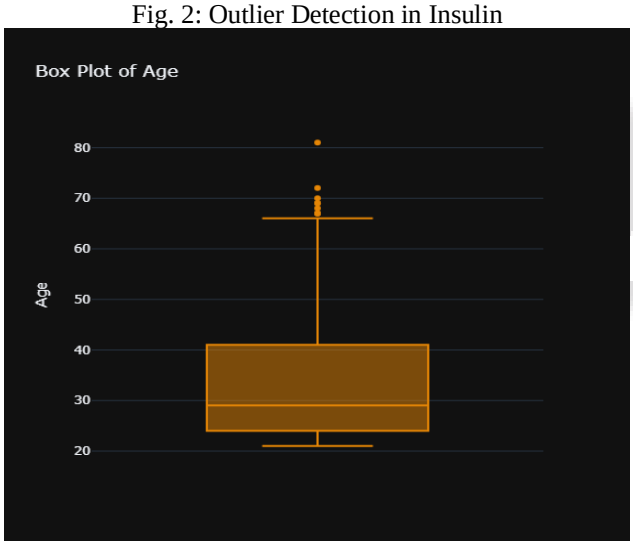

Fig. 3: Outlier Detection for Age

Outliers were detected for all the features using boxplot. As the part of preprocessing these outliers were removed in the dataset. The performance was measured before the outlier and after the outlier removal. A significance difference was seen in this study.

V. Results and Discussion

The application of machine learning to large datasets provides significant diabetes prediction abilities by finding precise analytic patterns beyond traditional medical diagnosis techniques. The enhanced forecasting accuracy lets medical practitioners base their choices on more knowledgeable information. The collected approach implemented by Random Forest led to its most effective performance compared to all investigated models due to its reduced overfitting and minimized variance effect.

The boost problem-solving technique of weak classifiers through repeated improvement allowed AdaBoost to excel in situations where classification remained troublesome. Naive Bayes joined Logistic Regression as simpler models which provided weak detection of nonlinear patterns in the data regardless of their relative explainability. The predicting accuracy of K-Nearest Neighbors (KNN) remains high yet this approach shows challenges when dealing with big datasets.

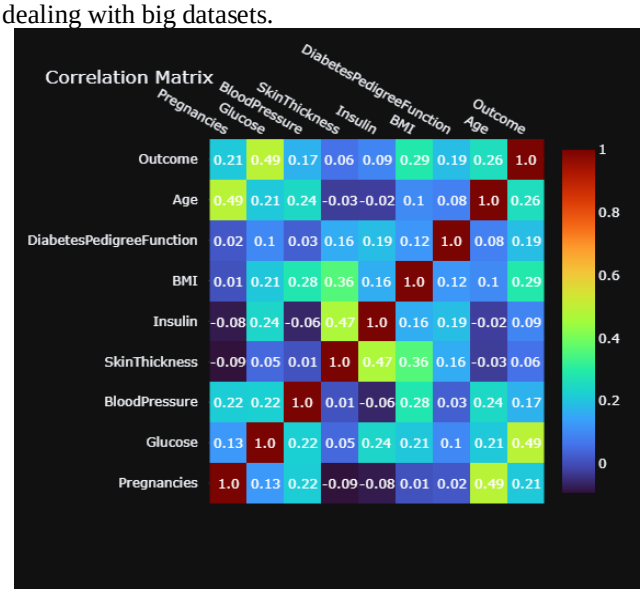

Fig. 4: Correlation matrix with a colorful heatmap

In addition to increasing diabetes early detection, machine learning allows for tailored risk assessments, which improves patient management. For instance, models can be customized to fit the unique features of each and every patient, enabling early lifestyle changes and pharmacological therapies to postpone growth of disease. By focusing attention on high-risk individuals, efficient resource allocation can be accomplished, lowering healthcare expenses and improving results. Tailored and suitable treatment regimens are also supported by predictive analytics, which improves patient involvement and attachment by providing real-time feedback. In the end, this improves quality of life and reduce problems.

This section illustrates the decision tree training. Predictions and prediction probabilities are made for the test set after fitting the model to the training data. The Receiver Operating Characteristic (ROC) curve is plotted to determine the performance of the model, illustrating the False Positive Rate (FPR) vs. True Positive Rate (TPR) trade-off.

VI. Future Scope

Uses of Machine learning in future offers improvements in accuracy and early detection, in diabetes prediction appear promise. If we want to improve predictive volume and enable more accurate risk assessments, one possible research avenue is to purify machine learning algorithms. By doing examining extensive and complex datasets, these types of models may be able to spot trends that ordinary techniques might overlook, which could lead to an earlier and more accurate diagnosis of diabetes. Some Advancements in models could incorporate different data sources, including genetic data and

electronic health records (EHR), which could greatly help preventative healthcare.

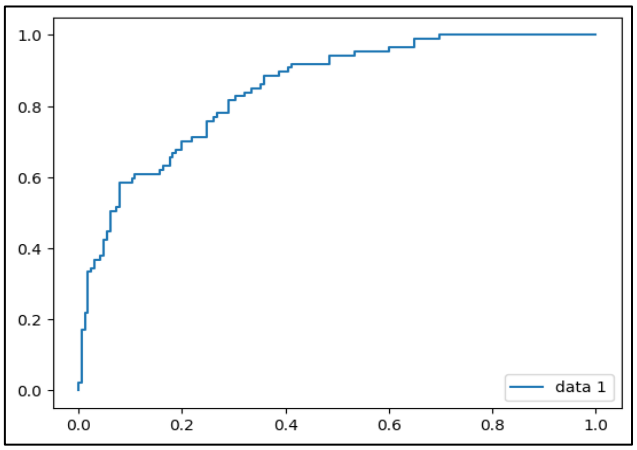

Fig. 5: ROC for decision tree

Another area of future attention will be Personalized risk assessment and management. The Ability for personalized care is expanded by machine learning's volume to customize forecasts based on genetic markers, lifestyle information, and individual health profiles. Only because of this personalization, medical professionals can provide suitable customized treatment and prevention strategies according to each patient's specified risk factors, which may improve patient outcomes. By calculating patient data and modifying therapies as necessary, machine learning (ML)-driven models may also enhance treatment plans in real-time and lowering negative side effects and enhancing the overall value of diabetes.

VII. Conclusion

Machine learning can help with improved disease management and the efficient use of healthcare resources, both of are essential for managing diabetes on a larger scale. By expecting possible issues, ML can support proactive management by allowing for prompt monitoring and intervention. As studies progress, machine learning's ability to lower diabetes-related complications, boost patient participation, and improve quality of life may change the face of diabetes care and make it more efficient and long-lasting. This paper presents the study on outlier and their impacts on performance of machine learning algorithms. This paper concludes that outlier affect the performance of the model. We achieved good accuracy for decision tree algorithm in the prediction of diabetes. Additionally, by helping in the identification of high-risk individuals, predictive models may enable healthcare systems to concentrate resources on those who require prompt intervention, lowering overall costs and enhancing the quality of care.

References

[1] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, et al., "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," Diabetes Res. Clin. Pract., vol. 183, p. 109119, 2022. https://doi.org/10.1016/j.diabres.2021.109119

[2] C. G. Awuchi, C. K. Echeta, and V. S. Igwe, "Diabetes and the Nutrition and Diets for Its Prevention and Treatment: A Systematic Review and Dietetic Perspective," Health Sciences Research, vol. 6, no. 1, pp. 5–19, 2020. [Online]. Available: http://aascit.org/journal/archive2?journalId=908&paperId=7753

[3] K. M. Rani, "Diabetes Prediction Using Machine Learning," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 6, no. 4, pp. 294–305, 2020. [Online]. Available: http://ijsrcseit.com/

[4] S. A. Tanim, A. R. Aurnob, T. E. Shrestha, M. R. I. Emon, M. F. Mridha, and M. S. U. Miah, “Explainable deep learning for diabetes diagnosis with DeepNetX2,” Biomedical Signal Processing and Control, vol. 99, p. 106902, 2025.

[5] A. Awasthi, "Diabetes Prediction using Machine Learning: A Review," International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 5, pp. 5146–5152, 2022. [Online]. Available: http://www.ijraset.com/

[6] M. Panda, D. Mishra, S. Patro, and S. S. Reddy, "Prediction of diabetes disease using machine learning algorithms," IAES Int. J. Artif. Intell. (IJ-AI), vol. 11, no. 1, pp. 284–290, 2022. [Online]. Available: https://ijai.iaescore.com/

[7] A. A. Abaker and F. A. Saeed, "A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications," Informatica, vol. 45, no. 1, 2021. [Online]. Available: https://informatica.si/

[8] V. Sapra and M. L. Saini, “Deep learning network for identification of ischemia using clinical data,” Int. J. Eng. Adv. Technol., vol. 8, no. 5, pp. 2357–2363, 2019

[9] M. L. Saini, B. Tripathi, J. Kaushal, and A. Garg, “A Hybrid Model for Diagnosis Cardiovascular Disease Using Clinical Features, ECG and MRI,” IEEE 1st Int. Conf. Ambient Intell. Knowl. Informatics Ind. Electron. AIKIIE 2023, 2023, doi: 10.1109/AIKIIE60097.2023.10390299

[10] K. Bansal, M. L. Saini, Rahul, K. Bhardwaj, and L. Prajapati, “Acne Skin Disease Detection Using Convolutional Neural Network Model,” Proc. - Int. Conf. Technol. Adv. Comput. Sci. ICTACS 2023, pp. 249–255, 2023, doi: 10.1109/ICTACS59847.2023.10389831.

[11] A. Jain, M. Lal Saini, A. Saklani, and A. Biju, “Tinea-Corporis Skin Disease Detection Using CNN and Kernel SVM,” Proc. - Int. Conf. Technol. Adv. Comput. Sci. ICTACS 2023, pp. 157–161, 2023, doi: 10.1109/ICTACS59847.2023.10389866

[12] E. G. Kumar, M. Lal Saini, S. A. Khadar Ali, and B. B. Teja, “A Clinical Support System for Prediction of Heart Disease using Ensemble Learning Techniques,” Int. Conf. Sustain. Commun. Networks Appl. ICSCNA 2023 - Proc., pp. 926–931, 2023, doi: 10.1109/ICSCNA58489.2023.10370569.

[13] S. P. Kumar Mygapula, M. Lal Saini, and C. S. Raj Dheeraj, “Performance Evaluation of Machine Learning Algorithms for Prediction of Cardiac Failure,” Int. Conf.

Sustain. Commun. Networks Appl. ICSCNA 2023 - Proc., pp. 1599–1604, 2023, doi: 10.1109/ICSCNA58489.2023.10368606

[14] Y. Singh, M. Saini, and Savita, "Impact and Performance Analysis of Various Activation Functions for Classification Problems," Proc. IEEE InC4 2023 - 2023 IEEE Int. Conf. Contemp. Comput. Commun., 2023, doi: 10.1109/InC457730.2023.10263129

[15] V. Sapra and M. L. Saini, "Computational intelligence for detection of coronary artery disease with optimized features," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 6, pp. 144–148, 2019

[16] V. Sapra, M. Saini, and L. Verma, "Identification of Coronary Artery Disease using Artificial Neural Network and Case-Based Reasoning," Recent Adv. Comput. Sci. Commun., vol. 14, no. 8, pp. 2651–2661, 2020, doi: 10.2174/2666255813999200613225404

[17] Tanushree Bharti; Yogam Singh; Mudit Jain; Ankita Kumari. "Improving Quality of Medical Scans using GANs." Volume. 9 Issue.12, December-2024 International Journal of Innovative Science and Research Technology (IJISRT), 1125-1131, https://doi.org/10.5281/zenodo.14557003

[18] Rawal, P., Ahuja, D., Saini, M. L., Chandel, H., & Raj, R. (2024). Cancer Detection and Treatment Using Explainable AI. International Journal of Scientific Research and Modern Technology, 3(9). https://doi.org/10.38124/ijsrmt.v3i9.47

[19] Barik, M. B., & Sidharth. (2024). Removing Unwanted Noise from Real Scene Images using GANs. International Journal of Scientific Research and Modern Technology, 3(10). https://doi.org/10.38124/ijsrmt.v3i10.60

