



MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING

Vidhi¹, Arpita², Anshika³, Nancy⁴, Ms. Ritu Rani⁵

Department of CSE-AI&DS,
Panipat Institute of Engineering & Technology, Samalkha, India^{1,2,3,4,5}
⁵Mail ID: ritu.cse.ai-ds@piet.co.in

Abstract--*This paper presents a unified machine-learning-based framework for preliminary prediction of heart disease, diabetes, and facial skin diseases. The system uses supervised classifiers for structured clinical data and a convolutional neural network (CNN) for image-based skin-disease classification, followed by integration through a Streamlit web interface. The project report describes preprocessing, feature engineering, model selection, hyperparameter tuning, evaluation, and report generation as core stages of the workflow. The reported CNN experiment achieved 91.40% training accuracy, 63.78% validation accuracy, and 62.70% test accuracy after 10 epochs; the gap between training and validation/test performance indicates a need for stronger generalization. Numerical accuracy values for the heart- and diabetes-prediction models are not explicitly reported in the source project report and are therefore not fabricated here. The framework is intended as a decision-support and screening aid rather than a replacement for clinical diagnosis.*

Keywords: Multiple disease prediction, machine learning, deep learning, CNN, Streamlit, healthcare AI.

INTRODUCTION

The increasing availability of healthcare data has created opportunities for machine learning (ML) systems that can assist with early disease-risk assessment. Prior work has explored supervised learning, feature selection, ensemble methods, and deep learning for disease prediction, while comparative studies have shown that model choice can substantially affect performance across medical datasets [1], [4], [6]. The project under study extends this idea to a unified platform covering three heterogeneous prediction tasks: heart disease and diabetes from structured clinical variables, and facial skin disease from images. The source report describes the goal as providing fast, accessible, data-driven preliminary predictions and a simple user interface for clinical-data entry and image upload [1], [2].

The main contribution of the implemented system is architectural rather than a single new learning algorithm: different models are selected according to the data modality and then exposed through one application. Logistic Regression, Random Forest, and Support Vector Machine (SVM) are considered for tabular prediction, while a CNN is used for image classification [2], [3]. The framework also incorporates preprocessing, feature engineering, cross-validation, hyperparameter tuning, and Streamlit-based deployment [2], [4].

The objectives are to: (i) develop predictive models for heart disease and diabetes from clinical/lifestyle attributes; (ii) classify facial skin-disease images using a CNN; and (iii) integrate the models into a unified web interface that returns predictions and downloadable reports [1], [2].



LITERATURE REVIEW

Research on automated disease prediction has progressed from individual binary classifiers toward systems that combine several prediction tasks and data types. Sana reported a multi-disease ML system with emphasis on feature selection, while Venkatesh combined ML and deep learning with a Streamlit interface [1], [2]. Bharath C. et al. compared decision-tree, Random Forest, and SVM approaches to identify suitable models for different diseases [3]. Ramesh et al. investigated feasible multi-disease prediction with attention to computational efficiency and scalability [4]. Broader studies have also compared supervised learning algorithms for medical prediction and highlighted the importance of dataset characteristics and evaluation methodology [5], [6].

Study	Approach	Key limitation / gap
[1] Sana (2024)	ML-based multi-disease prediction; feature selection.	Dataset imbalance; interpretability.
[2] Venkatesh (2023)	ML + deep learning + Streamlit.	Further accuracy/real-time validation needed.
[3] Bharath C. et al. (2024)	Decision Tree, Random Forest, SVM comparison.	Model comparison and feature selection remain dataset-dependent.
[4] Ramesh et al. (2023)	Scalable multi-disease ML prediction.	Dataset diversity and scalability.
[5], [6]	Disease-prediction survey and supervised ML comparison.	Generalization and clinical applicability remain important.

TABLE I. SUMMARY OF RELATED WORK [1]–[6]

The literature indicates three recurring issues relevant to the present implementation: heterogeneous data modalities, class imbalance/generalization, and interpretability. The source project addresses heterogeneity by using separate tabular and image models, and it identifies explainability and demographic generalization as future concerns [1], [2], [4].

PROBLEM STATEMENT AND SYSTEM OVERVIEW

Conventional diagnosis of diabetes, cardiovascular disease, and skin disorders can involve different measurements, laboratory tests, clinical examinations, or specialist review. The project therefore frames the computational problem as a first-level screening task in which minimal structured attributes or an image are processed by a corresponding predictive model [1], [2]. The system is not presented as a substitute for professional diagnosis; its role is to provide preliminary decision support.

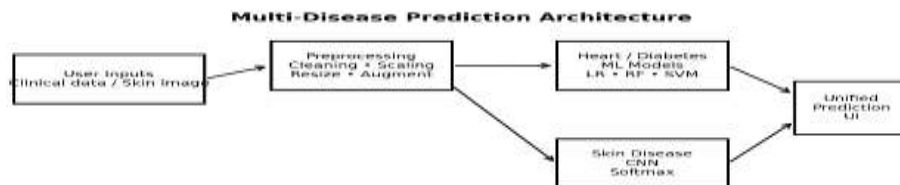


Fig. 1. Unified architecture for heterogeneous disease-prediction tasks.



METHODOLOGY

A. Data Sources and Input Features

The source report describes three datasets. Heart-disease data contain attributes such as age, sex, blood pressure, cholesterol, chest-pain type, heart rate, and ECG-related variables. The diabetes model uses the PIMA Indian Diabetes dataset with variables including pregnancies, glucose, BMI, insulin, and age. The image branch contains labeled dermatological images representing conditions such as acne, actinic keratosis, basal cell carcinoma, eczema, and rosacea [1], [2].

Task	Modality	Representative inputs	Model family
Heart disease	Tabular	Age, sex, BP, cholesterol, chest pain, ECG	LR / RF / SVM
Diabetes	Tabular	Pregnancy, glucose, BMI, insulin, age	LR / RF / SVM
Skin disease	Image	Facial/skin image	CNN + softmax

TABLE II. DATA MODALITIES AND MODEL ASSIGNMENT [1], [2]

B. Preprocessing and Feature Engineering

For structured data, preprocessing includes missing-value handling, outlier treatment, categorical encoding, and normalization/scaling. For images, resizing, normalization, rotation, flipping, and contrast-related augmentation are described. The report also proposes SMOTE or GAN-based augmentation for imbalance, and correlation analysis, PCA, or recursive feature elimination for feature selection [1], [2].

Development Methodology

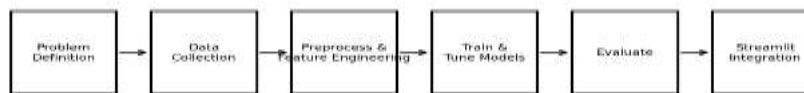


Fig. 2. Development workflow derived from the project methodology [1], [2].

C. Model Selection and Training

Logistic Regression is used as an interpretable baseline for binary classification; Random Forest provides an ensemble alternative with feature-importance capability; and SVM is suitable where decision boundaries are important [2], [3], [6]. For images, the project uses a CNN composed of convolutional layers for feature extraction, pooling for dimensionality reduction, dense layers for classification, and a softmax output for multi-class prediction [2]. Training uses train/validation/test separation, cross-validation, and hyperparameter tuning. GridSearchCV and regularization are specifically mentioned for improving generalization of tabular models [2].

D. Evaluation Metrics

The evaluation plan uses accuracy, precision, recall, F1-score, and ROC-AUC for the structured prediction tasks, with classification reports and confusion matrices for the image classifier [1], [2]. For a binary task, accuracy is defined as $(TP+TN)/(TP+TN+FP+FN)$, while precision and recall quantify the correctness and coverage of positive predictions. F1-score is the harmonic mean of precision and recall.



These metrics are preferred together because accuracy alone can be misleading under class imbalance [5], [6].

E. Application Integration and Privacy

The trained models are integrated into a Streamlit application. Users can select diabetes, heart disease, or skin disease prediction; enter the relevant clinical variables or upload an image; view the prediction; and generate a downloadable report [1], [2]. The project also identifies anonymization, encryption, informed consent, and attention to HIPAA/GDPR requirements as responsible-AI considerations [2].

RESULTS AND DISCUSSION

The project report demonstrates the end-to-end interface for diabetes, heart disease, and facial skin-disease prediction, including report generation. The report also contains an explicit CNN training log, which permits a numerical result to be reproduced without inventing missing performance values [1].

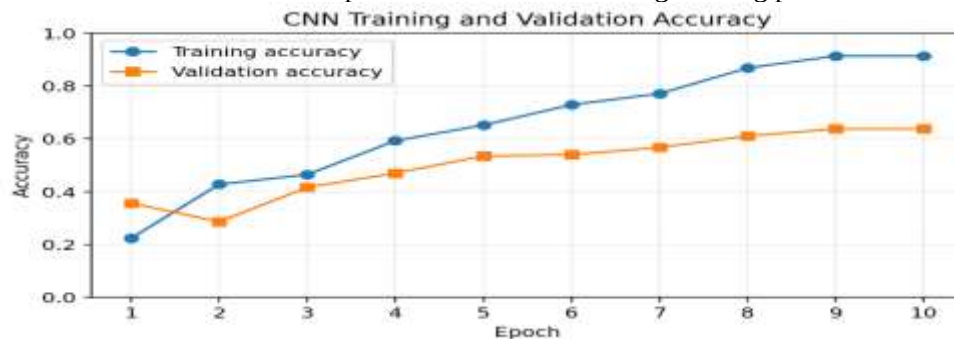


Fig. 3. CNN training/validation accuracy reproduced from the reported 10-epoch log [1].

Model / task	Training accuracy	Validation accuracy	Test accuracy
CNN skin-disease classifier	91.40%	63.78%	62.70%
Heart disease classifier	Not numerically reported	Not numerically reported	Not numerically reported
Diabetes classifier	Not numerically reported	Not numerically reported	Not numerically reported

TABLE III. REPORTED MODEL RESULTS [1]

The CNN training accuracy rises to 91.40%, while validation accuracy reaches 63.78% and the final test accuracy is 62.70% [1]. The approximately 28.7 percentage-point gap between training and validation accuracy is consistent with a generalization problem and suggests that additional regularization, stronger augmentation, transfer learning, class balancing, or a larger and more diverse image dataset should be investigated. These improvements are aligned with the project's identified need for larger and more diverse datasets and better generalization [1], [4]. Importantly, the source report does not provide numerical accuracy for the heart and diabetes models; reporting invented values would misrepresent the experiment.

The application-level results are also meaningful: the source report shows separate interfaces for diabetes, heart disease, and skin disease, together with a generated prediction report [1]. This supports the feasibility of the multi-model integration objective, although interface functionality should not be interpreted as clinical validation.



LIMITATIONS AND FUTURE DIRECTIONS

The principal limitations identified in the source material are dataset imbalance, limited diversity, image-quality variation, model explainability, and generalization across demographic groups. The reported CNN accuracy gap strengthens the case for addressing overfitting before real-world deployment. Future work described by the project includes more advanced deep-learning architectures, attention mechanisms, multimodal health data, wearable-device integration, personalized/adaptive prediction, telemedicine integration, explainable AI, mobile deployment, clinical validation, and continuous retraining. A practical research extension would be to benchmark the three tabular classifiers under the same stratified cross-validation protocol, report complete confusion matrices and ROC-AUC values, and then compare them with calibrated ensemble models. For the image branch, transfer learning and class-balanced evaluation could be used to reduce the observed training-validation gap. Such experiments would make the system more scientifically reproducible and clinically interpretable.

CONCLUSION

This paper reformulates the submitted project as an IEEE-style research manuscript describing a unified multiple disease prediction framework. The approach combines classical supervised ML for structured clinical data with a CNN for facial skin-disease images and exposes all three tasks through a Streamlit interface. The available experimental evidence shows that the CNN reached 91.40% training accuracy, 63.78% validation accuracy, and 62.70% test accuracy, demonstrating successful learning but also a substantial generalization gap. The system therefore provides a useful prototype for preliminary screening research, while stronger external validation, complete numerical reporting for all models, larger datasets, explainability, fairness analysis, and clinical collaboration are required before any clinical deployment.

REFERENCES

- [1] S. Sana, "Multi Disease Prediction System Using Machine Learning," *International Journal of Research Publication and Reviews*, vol. 6, no. 4, 2024.
- [2] M. Venkatesh, "Multiple Disease Prediction Using Machine Learning, Deep Learning and Stream-Lit," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 5, no. 7, 2023.
- [3] G. P. Deekshitha, M. P. Deepak, K. R. Gagan, K. S. Mohan Kumar, and Bharath C., "Multi Disease Prediction Using Machine Learning Algorithms," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 13, no. 6, pp. 11477–11484, Jun. 2024.
- [4] B. Ramesh, G. Srinivas, P. R. Praneeth Reddy, M. H. Rasool, D. Rawat, and M. Sundaray, "Feasible Prediction of Multiple Diseases using Machine Learning," *E3S Web of Conferences*, vol. 430, Art. no. 01051, 2023, doi: 10.1051/e3sconf/202343001051.
- [5] S. Kumar and A. Garg, "Predictive Analytics: A Survey of Machine Learning Techniques for Disease Prediction," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 6, pp. 312–317, 2018.
- [6] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing Different Supervised Machine Learning Algorithms for Disease Prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Art. no. 281, 2019, doi: 10.1186/s12911-019-1004-8.