Statistical Machine Learning vs. Deep Learning: Performance and Application in Healthcare

Sean Anderson
ECE Department
Stevens Institute of Technology
Hoboken, NJ, United States
sanders8@stevens.edu

Eli Shtindler
ECE Department
Stevens Institute of Technology
Hoboken, NJ, United States
eshtindl@stevens.edu

Christopher Spadavecchia
ECE Department
Stevens Institute of Technology
Hoboken, NJ, United States
cspadave@stevens.edu

***Abstract*—The goal of this project is to evaluate the effectiveness of different artificial intelligence models in the healthcare industry. To accomplish this, we will compare traditional machine learning models with deep learning models using both image and tabular datasets. The project incorporates a range of machine learning techniques, including Decision Trees, Random Forests, and Support Vector Machines, alongside neural network-based approaches. We will assess and compare these models using key performance metrics such as accuracy, precision, recall, and F1 score to determine which method offers the best results.**

I. Introduction

This project explores whether machine learning or deep learning models are more effective when applied to different types of datasets in the healthcare industry. As one of the most vital sectors globally, healthcare continues to adopt artificial intelligence to improve diagnostics, treatment, and patient care. Ensuring that AI models are optimized for this field is essential for maximizing their impact. To investigate this, we will compare traditional machine learning algorithms—such as Decision Trees, Random Forests, and Support Vector Machines—with deep learning models, specifically neural networks. The analysis will be conducted using two distinct datasets: a image dataset used to detect Pneumonia and a tabular dataset used to identify diabetes. By testing both types of models on image and structured data, we aim to determine which approach performs better in each context. Performance will be evaluated using key metrics including accuracy, precision, recall, and F1 score. The results of this study may inform future AI applications in healthcare by highlighting which techniques are best suited for particular data types and diagnostic tasks.

II. Related Work

Artificial intelligence has played a growing role in the healthcare industry, offering promising advancements in disease detection, medical decision support, and patient outcome prediction. A wide variety of AI approaches have been applied to medical datasets, ranging from traditional machine learning methods to more advanced deep learning architectures[1].

Machine learning algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Gradient Boosting have been extensively used to analyze structured tabular data like electronic health records, lab test results, and health screenings[2]. For example, Random Forest classifiers have been successfully applied to diabetes prediction, demonstrating strong performance due to their ability to handle missing data and reduce overfitting[3]. SVMs are also effective in binary classification problems in healthcare, especially on small datasets with many features[4]. Ensemble methods such as XGBoost and AdaBoost have gained popularity for their high predictive accuracy and computational efficiency[5]. These models are particularly useful in handling large and imbalanced datasets, which are common in healthcare, where identifying rare positive cases is critical. By combining multiple weak learners, ensemble models often outperform single algorithms[6].

In contrast, deep learning methods such as artificial neural networks (ANNs) and convolutional neural networks (CNNs) are especially well suited for unstructured data types like medical images, audio signals, and free-text clinical notes[7]. CNNs have become a leading choice for diagnostic imaging tasks, including tuberculosis, pneumonia, and COVID-19 detection from chest X-rays and CT scans[8]. These models automatically extract hierarchical features from raw pixels, reducing the need for manual preprocessing or feature selection. Other deep learning architectures, including recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), have been applied to time-series data in healthcare. This includes use cases such as ECG signal interpretation, glucose level tracking, and monitoring patient vitals in critical care settings[9]. More recently, transformer-based models have emerged as powerful tools for analyzing sequential clinical data and medical texts, offering strong performance in tasks that require contextual understanding and long-range dependency handling[10].

Numerous studies have compared traditional machine learning techniques with deep learning models across various healthcare applications. These comparisons indicate that model effectiveness often depends on the characteristics of the dataset. Machine learning models tend to perform well on smaller, structured datasets, while deep learning approaches typically excel in analyzing images and large, high-dimensional data sources[11].

III. Our Solution

A. Dataset Research

While researching for this project, we explored a variety of datasets and methodologies to find those best suited for comparing machine learning and deep learning models in

healthcare. Our goal was to identify datasets that would allow for a fair and meaningful comparison across different types of AI techniques, ideally representing both structured and unstructured data.

Initially, we experimented with the Jester Dataset, a large-scale video dataset developed by Qualcomm for gesture recognition tasks. Although rich in content and widely used in the field of computer vision, we quickly realized that it was not compatible with the machine learning models we intended to use. The dataset is tailored for spatiotemporal deep learning models like 3D CNNs or RNN-based architectures, which can process video frame sequences effectively. In contrast, traditional machine learning algorithms require structured, tabular input formats, and converting video data into a usable format proved to be resource-intensive and technically challenging.

For the tabular component of our project, we initially selected a general healthcare dataset that contained a wide range of patient-related information. The features in this dataset included patient name, age, gender, blood type, medical condition, date of admission, attending doctor, hospital name, insurance provider, and billing amount. Our objective was to use the categorical and numerical variables to predict the billing amount, which we designated as the target variable. Because billing amount is a continuous variable and most traditional machine learning models are better suited for classification tasks rather than regression in this context, we decided to convert the billing amounts into discrete categories through a process known as binning.

Our initial binning strategy divided the billing amounts into six categories in an attempt to capture a detailed spectrum of healthcare costs. This approach, however, quickly revealed several issues. First, the data distribution across the six bins was highly imbalanced, with a majority of the records falling into a few specific bins and very few examples in the others. This imbalance hindered the ability of our machine learning models to learn meaningful patterns and resulted in disproportionately low accuracy and poor generalization, especially on the underrepresented classes. Second, the features in the dataset appeared to offer little predictive value. For example, variables like hospital or doctor name may have had little relevance to cost in the absence of contextual data such as procedure codes, duration of stay, or severity of illness.

To address the imbalance and simplify the classification problem, we revised our binning strategy and reduced the number of categories from six to three. The new bins were defined as follows: low billing amounts (under $15,000), medium billing amounts ($15,000 to $35,000), and high billing amounts (above $35,000). This modification helped to distribute the data more evenly and offered a more manageable classification task for our models. We applied a variety of machine learning algorithms, including Decision Trees, Random Forests, and Support Vector Machines, to assess the dataset under this new structure. We also performed data preprocessing steps such as label encoding for categorical variables and normalization for numerical fields to prepare the dataset for training.

Despite these adjustments, the models continued to exhibit unsatisfactory performance across all major evaluation metrics, including accuracy, precision, recall, and F1 score. Further analysis revealed that many of the dataset's features lacked strong or consistent relationships with the billing amount. In other words, there was no clear signal in the input data that could be reliably associated with the output. For instance, variables like gender, blood type, or the name of the insurance provider did not show meaningful statistical correlation with how much a patient was charged. This suggested that either key explanatory variables were missing or the dataset was too noisy and arbitrary for our intended predictive task. These issues significantly limited the dataset's utility for training effective classification models.

Recognizing these limitations, we decided to pivot and search for datasets that were more structured, clinically relevant, and likely to exhibit stronger relationships between features and outcomes. Our goal was to find datasets that not only aligned better with the strengths of both machine learning and deep learning models, but also allowed us to make meaningful comparisons between the two approaches. After further research and evaluation, we identified two suitable datasets: the Pneumonia Prediction Dataset and the Diabetes Prediction Dataset.

B. Description of the Tabular Dataset

The Diabetes Prediction Dataset consists of multiple clinical and demographic features used to assess the likelihood of a patient being diagnosed with diabetes. In addition to the prediction variable, which indicates whether or not the patient has diabetes, the dataset includes eight key input variables: gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, and blood glucose level. These features were selected based on their established relevance to diabetes risk in clinical practice and provide a diverse set of indicators, ranging from lifestyle and demographic factors to physiological and biochemical metrics.

Before conducting any form of analysis or training predictive models, we carried out an essential preprocessing phase to ensure the dataset was clean, consistent, and ready for use. The first step involved removing duplicate records. Duplicate entries, if not addressed, can introduce bias, reinforce misleading patterns, and reduce the validity of analytical results. Eliminating these duplicates ensured that each data point represented a unique patient and that observed trends reflected true underlying relationships.

After cleaning the dataset, we performed exploratory data analysis to investigate possible correlations between the input features and the presence of diabetes. One of the most significant patterns we identified was the relationship between age and diabetes. We observed a clear positive correlation: as patient age increased, the likelihood of being diagnosed with diabetes also increased. This finding aligns with widely accepted medical knowledge that age is a major risk factor for type 2 diabetes. The increased risk in older populations can be attributed to a combination of metabolic changes, increased insulin resistance, and longer-term exposure to other contributing health conditions.

To illustrate this relationship, we created a line graph showing the frequency of diabetes diagnoses across various age ranges. The graph revealed a consistent upward trend, with the prevalence of diabetes rising notably in older age groups. This visualization supported our findings and emphasized

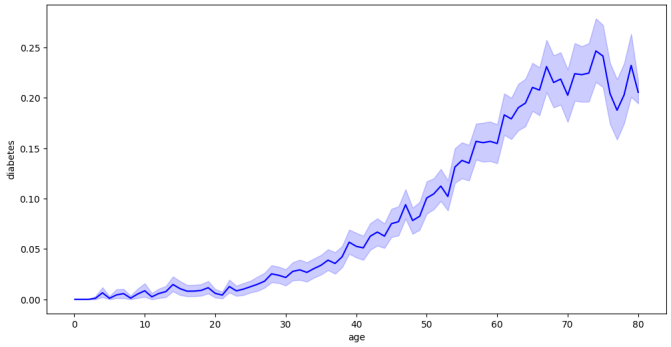

Figure 1. Age vs. Diabetes Plot

the importance of age as a predictive variable. The graph, shown in Figure 1, reinforced the observed correlation and highlighted age as a critical factor for further analysis and model development.

As part of the preprocessing stage, it was important to examine the data types of each variable in the dataset. Understanding the data types allowed us to determine which features required encoding before they could be used in model training. Some of the variables, such as hypertension, heart disease, and the diabetes outcome, were already represented as binary numerical values. These features used 0 to indicate absence and 1 to indicate presence of the condition, so no further transformation was needed for them.

However, other variables were categorical in nature. Specifically, the gender and smoking history columns were stored as string values. Since most artificial intelligence models require numerical input, these categorical variables needed to be converted into a usable format. To achieve this, we applied one-hot encoding, a technique that transforms categorical string data into a set of binary columns.

One-hot encoding works by creating a new binary column for each unique category in a feature. For example, the ”gender” feature, which contains categories such as ”Male” and ”Female”, is split into two new columns: one representing ”Male” and the other representing ”Female”. A value of 1 is placed under the corresponding category for each entry, while the other is set to 0. Similarly, the ”smoking history” feature, which contains multiple categories such as ”never”, ”former”, ”current”, and ”unknown”, is expanded into separate binary columns for each label. This process allows the model to interpret the categorical data without introducing false numerical relationships or assumptions.

The remaining variables in the dataset, including age, BMI, HbA1c level, and blood glucose level, were already stored as floating-point numerical values. These did not require encoding and were ready for use as-is. After completing the encoding of all necessary features, the dataset consisted entirely of numerical values and was fully prepared for the modeling phase. At this point, all preprocessing tasks had been completed, and the dataset was ready for training and evaluation.

C. Description of the Image Dataset

The Chest X-ray Pneumonia Dataset is a curated collection of grayscale medical images used to classify patients as either having pneumonia or being normal. It was sourced from a clinical environment and is hosted on Kaggle by Paul Mooney. The dataset contains chest radiograph (X-ray) images of pediatric patients, divided into three folders: train, test, and val. Each folder includes two subdirectories labeled NORMAL and PNEUMONIA, reflecting the ground truth diagnosis for each image. The pneumonia cases include both viral and bacterial infections, though they are not separated by subtype in this version of the dataset.

Figure 2. Images from the Pneumonia Dataset

This dataset is particularly useful for deep learning applications, especially convolutional neural networks (CNNs), because it provides a binary classification challenge on real-world medical image data. Each image varies in size and resolution, and there is inherent variation in brightness, contrast, and position due to differences in scanning equipment and patient posture.

The Chest X-ray Pneumonia Dataset is a curated collection of grayscale medical images used to classify patients as either having pneumonia or being normal. Figure 2 presents sample images that illustrate the types of chest X-rays included in the dataset. The dataset contains chest radiograph (X-ray) images of pediatric patients, divided into three folders: train, test, and val. Each folder includes two subdirectories labeled NORMAL and PNEUMONIA, reflecting the ground truth diagnosis for each image. The pneumonia cases include both viral and bacterial infections, though they are not separated by subtype in this version of the dataset.

This dataset is particularly useful for deep learning applications, especially convolutional neural networks (CNNs), because it provides a binary classification challenge on real-world medical image data. Each image varies in size and resolution, and there is inherent variation in brightness, contrast, and position due to differences in scanning equipment and patient posture.

D. Statistical Machine Learning Algorithm for Tabular Dataset

To evaluate the predictive potential of the Diabetes Prediction Dataset, we implemented and tested two primary machine learning algorithms: the Decision Tree Classifier and the Random Forest Classifier. Both models were chosen for their interpretability, effectiveness on structured data, and relevance to classification tasks in healthcare settings.

The dataset was first prepared through preprocessing steps, including the removal of duplicate entries and the application of one-hot encoding to convert categorical variables such as gender and smoking history into numerical format. The target variable was the presence or absence of diabetes, and all other variables were used as input features. After preprocessing, the dataset was split into training and testing subsets using an 80-20 split, with stratification to preserve the class distribution.

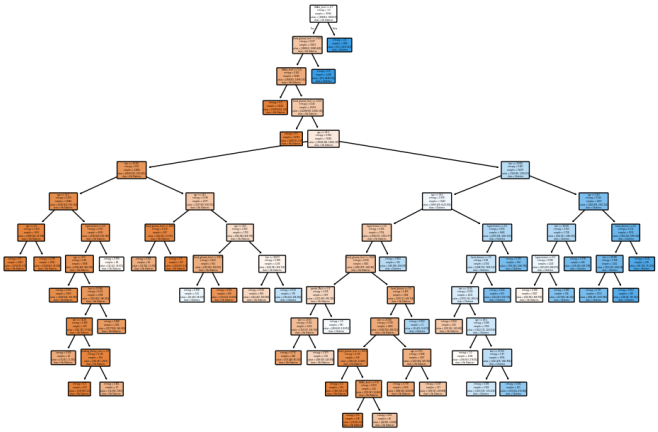

Figure 3. Decision Tree for the Diabetes Dataset

The first algorithm applied was the Decision Tree Classifier, a rule-based model that splits data based on feature values to make predictions. This model was trained using the entropy-based information gain as splitting criteria. Decision trees are particularly useful for understanding which variables contribute most to prediction, and they provide clear visual representations of decision paths. We analyzed the decision paths generated by the model and examined the importance of features, such as age and glucose levels, in determining diabetes status. he structure and decision paths of the trained Decision Tree model are illustrated in Figure 2, providing a visual representation of how the model classifies patients based on input features.

The second model tested was the Random Forest Classifier, which builds an ensemble of multiple decision trees and combines their outputs for improved accuracy and robustness. This model helps to reduce overfitting and enhances generalization by averaging the predictions of individual trees, each trained on a randomly sampled subset of the data. The Random Forest model consistently produced higher accuracy and more stable results across various performance metrics. Feature importance scores generated by the model also helped identify the most influential predictors of diabetes in the dataset.

E. Deep Learning Algorithm for Tabular Dataset

In continuing to evaluate the predictive potential of the Diabetes Prediction Dataset, a feedforward neural network was used to test the effectiveness of using a deep learning algorithm. This type of neural network was chosen since it could be simply implemented but was known to be well suited for classification tasks. It was also ideal for utilizing with this dataset since it can handle complex relationships including where an outcome may rely on multiple input features.

In preprocessing, before the training dataset could be used on the FNN, the categorical columns of the dataset, which includes gender and smoking history, were one hot encoded to allow the neural network to only have to work with numerical features. In order to have the FNN use less memory as well, numerical data values were scaled for optimization and training and testing input values were cast to float32.

During early testing, another issue seen was a lack of representation in the outcome by the ”have diabetes” class. This was due to a data balance issue since the ”does not have diabetes” class far outweighs the other. To fix this, the class weights of the two classes were rebalanced based on their sizes.

The FNN implementation includes 2 layers: the Hidden Layer and the Output Layer. In each epoch of training, the data was shuffled and a random batch of 128 sample subsets were chosen to train with. The model performed a forward pass and the predicted outputs were computed. Following, the loss was computed which accounted for class imbalance. Then, backpropagation was used to compute the gradients of the loss with respect to each weight and bias. These gradients were scaled by the learning rate of 0.01 and used to update the weights in a direction that would minimize the loss. Following 10 epochs, a random sample of data was used to report the current loss which helped monitor training. It was determined through multiple iterations of training that 25 epochs produced the best accuracy for the model.

Following training, the outcome of the testing data was predicted utilizing the model. This was done by establishing a threshold value that would be used to place the outcome in one of two classes. The outputs would fall between 0 and 1, and if the value was less than 0.5, it was placed in class 0, and if it was greater than 0.5, it was placed in class 1. The model was evaluated using accuracy, precision, recall, and f1-score of the testing dataset.

F. Statistical Machine Learning Algorithm for Image Dataset

While deep learning approaches are the standard for image classification, this project also explored the use of traditional statistical machine learning methods on the Chest X-ray Pneumonia Dataset. Specifically, a Support Vector Machine (SVM) classifier was applied to demonstrate how well a non-deep-learning model could perform on unstructured image data. Since most traditional algorithms require tabular input, a key part of this process involved transforming image data into a suitable format.

Each X-ray image, originally stored in grayscale, was resized to a fixed dimension of 224×224 pixels to maintain consistency across the dataset. These images were then flattened from 2D arrays into 1D vectors, effectively turning each image into a row of pixel intensity values. This transformation allowed the data to be interpreted as numerical features, similar to structured datasets used in classification problems. After reshaping, the pixel values were normalized to fall within the range [0, 1], helping the SVM model converge more efficiently.

Once the data was prepared, the SVM classifier was trained using a radial basis function (RBF) kernel, which is capable of capturing non-linear decision boundaries. Due to the high dimensionality of the input space (over 50,000 features per image), training time and memory usage were notable challenges. Additionally, since the dataset is imbalanced, with more pneumonia cases than normal images, class weights were adjusted in the training process to penalize misclassification of the minority class.

Evaluation was performed using accuracy, precision, recall, and F1 score. The SVM model achieved moderate accuracy,

correctly identifying many pneumonia cases, but it underperformed compared to deep learning approaches. This result highlighted the limitations of statistical models in image classification, particularly when spatial hierarchies and complex textures are involved. Nonetheless, the inclusion of an SVM provided valuable perspective on the importance of input representation and model complexity when working with medical imaging tasks.

G. Deep Learning Algorithm for Image Dataset

To leverage the full potential of unstructured medical image data, we implemented a deep learning approach using a Convolutional Neural Network (CNN) to classify chest X-rays as either showing signs of pneumonia or appearing normal. CNNs are particularly well suited for image-based tasks due to their ability to learn spatial hierarchies and visual features such as edges, textures, and shapes directly from pixel data.

The Chest X-ray Pneumonia Dataset was loaded and preprocessed using TensorFlow's image_dataset_from_directory() utility, which automatically labels and batches the images based on their folder structure. All images were resized to 224×224 pixels to ensure compatibility with standard CNN input dimensions. Pixel values were normalized to the [0, 1] range to improve training stability and convergence. The dataset was then divided into training, validation, and test sets, with a 10

The CNN architecture used in this project consisted of multiple convolutional layers, each followed by a ReLU activation function and max pooling. These layers enabled the network to extract increasingly complex features at deeper levels of the network. After several convolutional and pooling blocks, the output was flattened and passed through one or more fully connected (dense) layers. A final sigmoid activation function was used in the output layer to perform binary classification between pneumonia and normal cases.

To improve generalization and prevent overfitting, techniques such as dropout regularization and data augmentation (including random flips and rotations) were applied. The model was compiled with the binary cross-entropy loss function and optimized using the Adam optimizer, which provided adaptive learning rate adjustments during training. The network was trained over several epochs, with performance monitored on the validation set after each epoch.

The CNN achieved significantly better performance than traditional machine learning approaches, particularly in recall and F1 score—key metrics in a healthcare context where false negatives can lead to undiagnosed cases. This result underscores the effectiveness of deep learning for medical imaging tasks and demonstrates how CNNs can automatically discover visual patterns relevant for clinical diagnosis without requiring handcrafted features.

H. Implementation Details

After training and evaluating all four models, we obtained results that allowed us to compare machine learning and deep learning approaches across both image-based and tabular healthcare datasets. Model performance was assessed using four key evaluation metrics: accuracy, precision, recall, and F1 score. These metrics were derived from each model's corresponding confusion matrix, which breaks down the predictions into four categories: true positives, true negatives, false positives, and false negatives.

In the context of healthcare, understanding the implications of these outcomes is critical. False negatives, where a condition goes undetected, pose a significant risk because they may prevent a patient from receiving timely and potentially life-saving treatment. On the other hand, false positives, although less immediately dangerous, can still lead to unnecessary stress, additional testing, and the misuse of healthcare resources. While both types of errors are undesirable, the consequences of false negatives are typically more severe in clinical applications. For this reason, recall and F1 score, which emphasize the correct identification of positive cases, are especially important when evaluating model effectiveness in medical decision-making.

The performance results for tabular Diabetes Prediction Set can be seen in Figures 4, 5, and 6.

Metric	Score
Accuracy	0.89
Precision	0.94
Recall	0.89
F1 Score	0.91

Figure 4. Performance metrics for Decision Tree Classifier on the Diabetes Prediction Dataset.

Metric	Score
Accuracy	0.97
Precision	0.97
Recall	0.97
F1 Score	0.97

Figure 5. Performance metrics for the Random Forest Classifier on the Diabetes Prediction Dataset.

Metric	Score
Accuracy	0.89
Precision	0.94
Recall	0.89
F1 Score	0.90

Figure 6. Performance metrics for the Fully Neural Network on the Diabetes Prediction Dataset.

The performance results for the image Pneumonia Prediction Dataset can be seen in Figures 7 and 8.

Metric	Score
Accuracy	0.65
Precision	0.81
Recall	0.65
F1 Score	0.64

Figure 7. Performance metrics for the SVM on the Pneumonia Prediction Dataset.

Metric	Score
Accuracy	0.81
Precision	0.85
Recall	0.81
F1 Score	0.79

Figure 8. Performance metrics for the CNN on the Diabetes Prediction Dataset.

IV. Comparison

At the beginning of the project, we made predictions about which models would perform best on each type of dataset. We anticipated that traditional machine learning algorithms would yield better results on the tabular dataset, while neural networks would outperform other models on the image-based dataset.

After reviewing the performance results for each model on each dataset, we can confirm that our initial hypothesis was correct. The tabular Diabetes Prediction Dataset showed the strongest results with traditional machine learning models. The Random Forest Classifier achieved the highest overall performance, with an accuracy, precision, recall, and F1 score of 0.97, as shown in Figure 5. The Decision Tree Classifier and the Fully Connected Neural Network produced comparable results, each reaching an accuracy of 0.89 and F1 scores of 0.91 and 0.90 respectively (Figures 3 and 5). While the neural network performed well, it did not surpass the Random Forest, reinforcing the suitability of ensemble-based machine learning methods for structured, tabular data.

In contrast, the Pneumonia Prediction Dataset, which consists of chest X-ray images, revealed a clear advantage for deep learning approaches. The Convolutional Neural Network (CNN) outperformed the Support Vector Machine (SVM) across all four evaluation metrics. The CNN achieved an accuracy of 0.81 and an F1 score of 0.79 (Figure 7), compared to the SVM's accuracy of 0.65 and F1 score of 0.64 (Figure 6). This result reflects the CNN's ability to automatically learn and extract meaningful spatial features from image data, which traditional machine learning models like SVMs are not well-equipped to handle without extensive feature engineering.

Overall, the results support the conclusion that model effectiveness depends heavily on the nature of the dataset. Traditional machine learning algorithms, especially tree-based models, excel with structured numerical and categorical data, while deep learning models such as CNNs are more effective when dealing with unstructured image data. These findings emphasize the importance of aligning model architecture with the structure of the input data to maximize predictive performance in real-world healthcare applications.

V. Future Directions

This project provides a foundational comparison between traditional machine learning and deep learning models applied to both structured and unstructured healthcare datasets. While the results confirmed our initial hypothesis that machine learning performed best on tabular data and deep learning excelled on image-based data, there are still many opportunities for extending this work and improving its real-world applicability.

One important future direction involves incorporating more diverse and clinically complex datasets. For example, using multi-class classification problems, time-series data from continuous patient monitoring such as ECG or glucose levels, or multimodal datasets that combine imaging, lab results, and clinical notes would offer a more comprehensive view of patient health. These types of data would also provide more rigorous testing for model robustness in realistic clinical scenarios[15]. Datasets sourced from electronic health records are particularly valuable because they include temporal patterns, missing values, and longitudinal records that better reflect the complexity of healthcare environments.

In terms of modeling, this project could be significantly enhanced by implementing more advanced deep learning architectures. While a basic convolutional neural network was used for image classification, pretrained models such as ResNet, DenseNet, or InceptionV3 could improve performance when applied using transfer learning. These models are optimized to extract high-level spatial features and have been shown to outperform custom CNNs in many diagnostic imaging tasks[14]. For tabular data, ensemble models like XGBoost or LightGBM could offer additional improvements by combining the strengths of gradient boosting with the structure of traditional pipelines.

Interpretability is another critical area to explore. In healthcare, trust in AI predictions depends on understanding how those predictions are made. Tools like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) can help clinicians see which features most influenced a model's decision[12]. Including these tools in future iterations would increase transparency and support clinical validation, especially in settings where accountability and ethical considerations are essential.

Future work could also assess model performance across different patient populations, including age groups, socioeconomic backgrounds, or racial and ethnic demographics. Evaluating fairness and bias across subgroups can help ensure that AI systems do not inadvertently reinforce healthcare disparities[13]. Using fairness metrics and subgroup analysis can guide improvements in model design to promote equity in medical outcomes.

Another promising direction involves deployment of the models into a working prototype. Developing a web-based or mobile decision support tool would allow real-time predictions using new patient data. This would help evaluate usability, responsiveness, and workflow integration in clinical environments. Feedback from clinicians during this phase would also support iterative improvements and highlight practical barriers to adoption.

Lastly, working alongside healthcare professionals during all stages of model development, especially in feature selection, data interpretation, and performance evaluation, would greatly improve the clinical relevance of the final tool. Their expertise can guide which features are meaningful in practice, and their involvement can ensure the tool addresses genuine diagnostic and operational needs.

In conclusion, while this project provides strong evidence for the effectiveness of model types relative to data structure, it also lays the groundwork for continued exploration in real-world clinical AI. By expanding to richer datasets, applying more sophisticated models, focusing on interpretability and fairness, and progressing toward deployment, future research can help bring AI tools from the lab into everyday healthcare practice.

VI. Conclusion

This project aimed to evaluate and compare the performance of machine learning and deep learning models in the healthcare industry by applying them to two distinct types of datasets: a tabular dataset for diabetes prediction and an image-based dataset for pneumonia detection. The goal was to determine whether the structure and nature of the input data influence which type of model performs best. Through a structured process of data preprocessing, model training, and performance evaluation using standard metrics such as accuracy, precision, recall, and F1 score, we were able to develop and assess a range of models, including Decision Trees, Random Forests, Support Vector Machines (SVMs), and neural networks.

Our results confirmed the hypothesis that traditional machine learning algorithms are better suited for structured tabular data, while deep learning models excel when applied to unstructured image data. The Random Forest Classifier outperformed all other models on the diabetes dataset, achieving near-perfect performance metrics across the board. This reinforces the effectiveness of ensemble-based machine learning models in handling clinical datasets that include numerical and categorical variables. Conversely, the Convolutional Neural Network demonstrated superior performance on the pneumonia X-ray dataset, surpassing the SVM model by a significant margin. This result aligns with existing literature and practical experience in the field, as CNNs are specifically designed to process spatial features in image data and can automatically learn patterns that would be difficult to engineer manually.

Beyond model accuracy, this project also highlighted important considerations for implementing AI in healthcare. Preprocessing steps such as duplicate removal, one-hot encoding, and normalization were crucial for ensuring model compatibility and improving training outcomes. The challenges encountered with our initial datasets also emphasized the importance of choosing data that contain meaningful, clinically relevant features. Furthermore, the role of model interpretability and fairness was acknowledged as essential for real-world adoption, especially in high-stakes environments like medicine where transparency and trust are critical.

This study serves as a foundational comparison that not only validates theoretical expectations but also provides practical insights for future work. As artificial intelligence continues to evolve, its integration into healthcare systems will require ongoing assessment of which models best serve specific data types and clinical goals. Future research could explore more advanced architectures, incorporate richer and more diverse datasets, and prioritize interpretability tools such as SHAP or LIME to make model decisions more transparent to practitioners. Additionally, real-world deployment through prototypes and clinical feedback loops could bridge the gap between experimental accuracy and real-life utility.

In conclusion, this project demonstrates that selecting the right model for the right type of data is critical to achieving strong and reliable performance in healthcare AI applications. The findings provide evidence-based guidance for future efforts to build AI tools that can assist in diagnosis, risk prediction, and decision support. With continued research and collaboration between data scientists and healthcare professionals, these models have the potential to significantly enhance patient outcomes, optimize clinical workflows, and support the broader transformation of the medical field through intelligent automation.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, 2017.

[2] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, 2015.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and G. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, 2017.

[4] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12, 2006.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[6] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, 2007.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[9] O. Faust, Y. Hagiwara, T. J. Hong, M. H. L. Oh, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer Methods and Programs in Biomedicine*, vol. 161, 2018.

[10] I. Li, M. Bean, and X. Zhu, "BEHRT: Transformer for electronic health records," *Scientific Reports*, vol. 10, no. 1, 2020.

[11] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, J. Mottaghi, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, 2019.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, 2019.

[14] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, 2016.

[15] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, and others, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[16] Mustafa Tariq, "Diabetes Prediction Dataset," *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[17] P. T. Mooney, "Chest X-Ray Images (Pneumonia)," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[18] Prasad Narayana, “Healthcare Dataset,” *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/prasad22/healthcare-dataset

[19] Qualcomm AI Research, “Jester: A Large-Scale Human Action Video Dataset,” *Qualcomm Developer Network*, 2019. [Online]. Available: https://www.qualcomm.com/developer/software/jester-dataset.