**Literature Review**

The use of machine learning (ML) methods to predict diabetes has become increasingly popular in recent years, and many studies have contributed to this discipline. In order, to improve the accuracy and efficiency of diabetes diagnosis and treatment, various machine learning and deep learning techniques have been investigated in these studies, from traditional methods to AI-driven projects. This study presents timely findings that contribute to the expanding field of diabetes prediction using machine learning, highlighting significant advancements and illustrating the ongoing efforts to refine these technologies.

**Tejas N. Joshi and Prof. Pramila M. Chawan’s 2018 study, “Diabetes Detection Using Machine Learning,”** was one of the first studies in this area. The study used three supervised machine learning algorithms: support vector machine (SVM), logistic regression (LR), and artificial neural network (ANN) to identify patients with and without diabetes. The algorithms are based on key features such as blood pressure, skinfold thickness, insulin levels, BMI, age, diabetes, and diabetes spectrum function. SVM is particularly good at binary classification tasks due to its ability to produce optimal hyperplanes, but LR produces easy-to-understand results for binary outcome prediction. Since neural networks follow the patterns of the brain and have additional complexity, combining these patterns can increase accuracy. This study demonstrates the transformative potential of machine learning for diabetes research by supporting clinicians in early diagnosis and decision-making.

Following this study, **Bello A. Bodinga et al. conducted a large-scale study titled "Evaluation of some machine learning algorithms for diabetes prediction.".** This study used three supervised machine learning algorithms on the PIMA India diabetes dataset: logistic regression, random forest (RF), and decision tree (DT). This study highlights the importance of early diagnosis for effective diabetes management, where retrieval has the highest accuracy of 76% and RF comes second with 75%. The results show the best performance of logistic regression in handling the binary distribution function while also validating the robustness of random forests to noise and their ability to represent relationships. Accuracy and reliability of the prediction model. Used to predict blood glucose. Similarly, as shown by **Patil and Tamane (2018)**, SVMs consistently outperform other classifiers across datasets. **Komi et al. (2017)** analysed the risk of diabetes using data mining techniques and demonstrated the potential of complex analysis methods in clinical applications.

In the study titled **"Use of Different Machine Learning Techniques" by Subrata Kumer Paul et al.**, DT, K-nearest neighbours (KNN), Naive Bayes (NB), and RF are applied to the PIMA dataset after a series of intensive preprocessing steps, including incomplete data handling and standardization. RF showed the best performance with 86% accuracy and was praised for its performance in noise and missing data. To improve healthcare applications, this study proposes the combination of machine learning models with real-time data collection from IoT sensors.

**J. Dinesh, 2021**, signalling a move toward deep learning in diabetes prediction. The study used continuous oscillation deep neural networks to reduce overfitting and improve the prediction of blood glucose outcomes.

Previous studies by **Meng et al. (2013)** and **Bashir et al. (2016)** found that multi-layered methods outperform traditional ML methods when dealing with complex data, as suggested by the authors.

An article by **Saxena et al**. showed significant results obtained by comparing five groups, namely Naïve Bayes, random forest, logistic regression, neural network, and SVM, using the PIMA dataset. Logistic regression showed the best performance with 77.2% accuracy, which was effective in classifying binary tasks and showed that preliminary procedures such as process control are important.

**"For Prediction of Diabetes Mellitus" by Selma Aouamria et al.** showed improved prediction; this study combined soft selection techniques with three deep learning techniques: LSTM, deep neural networks (DNN), and convolutional neural networks (CNN). This method addresses the important health issues that come with diabetes. To overcome the problems caused by small datasets, the researchers use data fusion methods. Their model performed well on the PIMA India Diabetes Dataset (PIDD), achieved 98% accuracy on the Frankfurt Hospital German Diabetes Dataset (FHGDD), and 99.81% accuracy on the mixed dataset. In addition to addressing the issue of large datasets and reducing human error, these results also demonstrated the effectiveness of the model in predicting diabetes, as it was found to outperform the α distribution.

The research paper **"A Machine Learning Technique for Detection of Diabetes Mellitus" by O. M. Awoniran et al (2022)**, explores the application of machine learning algorithms for predicting diabetes mellitus, focusing on enhancing prediction accuracy through data pre-processing and the use of SVM, RFC, and DNN algorithms. The study uses data from the National Institute of Diabetes and Digestive and Kidney Diseases, pre-processed with dummy variables and PCA, achieving the highest accuracy of 89% with DNN, outperforming SVM and RFC. The literature review discusses various studies, including **Mujumdar and Vaidehi (2019)** on external factors in pipeline models, **Mitushi and Varma (2020)** on multiple classifiers with Gradient Boosting, and **Neha and Shruti (2020)** on Type-II diabetes prediction with Random Forest achieving 94% accuracy. Despite highlighting the importance of data pre-processing, the paper identifies gaps such as the limited exploration of ensemble methods and reliance on a single dataset, suggesting the need for diverse datasets and additional features to optimize prediction accuracy.

To address the usage of an ensemble model, the paper **"Prediction of Diabetes with its Symptoms Based on Machine Learning" by X. Xu, X. Huang, J. Ma, and X. Luo(2021)** highlights the superior performance of Random Forest in predicting diabetes, emphasizing its robustness in managing complex healthcare data. Ensemble models like Random Forest combine multiple decision trees to improve prediction accuracy and reduce the risk of overfitting. The research builds on previous work by scholars such as **Tripathi et al.(2020)**, who achieved an accuracy of 87.66% with Random Forest, and **Sisodia et al.(2020)**, who highlighted Naïve Bayes with an accuracy of 76.30%. Other studies by **Zou et al.(2019)** and **Butt et al.(2021)**. further validate the significance of ML algorithms in healthcare applications. The authors conducted rigorous preprocessing, including normalization, handling class imbalance, and feature importance evaluation, and used metrics like precision, recall, F1-score, and accuracy to assess model performance. The study underscores the efficacy of ensemble models, particularly Random Forest, in addressing the complexities of diabetes prediction and enhancing the predictive power of ML models in medical diagnostics.

**Prediction of Diabetes Mellitus using ML Techniques : A Systematic Overview” by T. Krishna Manaswini et al (2023)** examines various ML approaches applied to diabetic datasets to aid in the early diagnosis and management of Diabetes Mellitus. The paper highlights the use of ML techniques with Big Data Analytics tools such as Hadoop and MapReduce, as well as classifiers like Naïve Bayes, Decision Trees, SVM, KNN, Random Forest, and Gradient Boosting.

**"Technology for diabetes prediction" by Omoora et al.** demonstrated that XGBoost performs better overall because it can also handle nonlinear interactions. To provide personalized medical advice, the authors propose to create an AI system that combines cognitive processing with a user interface.

Research by **Devansh Sehgal et al., 2024, "AI-enabled Early Diabetes Prediction,"** demonstrated the various machine learning models for early diabetes diagnosis, including KNN, SVM, Gradient Boosting, Naïve Bayes, and LR. KNN is the best-performing algorithm with 75% accuracy and has proven its usefulness when used with Flask for real-time prediction. This study addresses issues such as data quality and sampling bias while demonstrating the potential of AI to transform healthcare by providing insights into diabetes risk.

Finally, the **study “Research and Comparison of Machine Learning Techniques for Blood Glucose Monitoring” by Kumar et al., 2024**. Five classifications were examined, namely LR, SVM, DT, RF, and KNN. The study found that RF achieved 92.23% accuracy. This study found that RF was able to predict negative interactions in relationships well, while LR performed poorly (74%). The authors emphasize that advanced techniques like RF are needed in complex medical cases. The study by **Meng et al.** compared many data mining models, such as random forests and SVM, to predict the results of the study by **Bashir et al.** The performance of multi-method learning algorithms such as KNN, NB, and SVM has been well investigated by **Kavakiotis et al.** in artificial intelligence with mobile health applications; comparison by **Rawat et al.** highlighted the importance of specific selection for improving patient outcomes, ranging from techniques such as transformation, logistic regression, and support vector machines (SVM) to more complex methods such as random forests and XGBoost.

The research paper titled **"A Comprehensive and Comparative Examination of Machine Learning Techniques for Diabetes Mellitus Prediction" by A. Kumar, A. S. Gill, J. P. Singh, and D. Ghosh (2024)** investigates the application of machine learning (ML) classifiers for predicting diabetes mellitus. The study employs five ML models—Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and k-Nearest Neighbors (k-NN)—to evaluate their effectiveness in diabetes prognosis. The RF classifier achieved the highest accuracy of 92.23%, demonstrating its strength in modeling complex, non-linear relationships, while LR showed a lower accuracy of 74.42%. The study builds upon previous work, such as **Mujumdar and Vaidehi (2019)** using LR for 96% accuracy, **Bhat et al. (2020)** applying RF with 98% accuracy, and **Hassan et al. (2021)** and **Jackins et al. (2022)** highlighting the efficacy of ensemble methods and SVM, with SVM reaching 96.54% on the dataset. The study emphasizes the potential of ML in reducing the workload of medical practitioners and enabling early diagnosis through comprehensive data preprocessing, including handling missing values, feature scaling, and correlation adjustments. The robust evaluation framework, including accuracy, precision, recall, and F1-score metrics, identified RF as the most effective model. The paper concludes by advocating for the broader integration of ML in healthcare, suggesting future research directions like clinical decision support systems to improve early disease detection and management.

Deep learning algorithms have improved prediction efficiency by managing large amounts of data and using techniques such as regular interpolation to avoid overfitting issues. In practice, it is shown how health can be improved by facilitating self-management and early diagnosis with the use of complex models such as the integration of deep learning. Effective diagnostic tools for diabetes must address issues such as sampling bias, computational costs, data quality, and ethical issues before they become available.

This literature review highlights the importance of developing machine learning methods in healthcare to support the timely management of diabetes and support new developments in predictive analytics in medicine.