Group 109 ML Project Final Report

Introduction/Background

Alzheimer's is a form of dementia that affects memory, thinking, and behavior in humans. Neural disruption in the brain causes gradual declines in cognitive abilities before eventual cell death. Our project's aim is to investigate how ML can address challenges in Alzheimer detection and management.

Existing literature already demonstrates ML algorithms' value in predicting Alzheimer onset; in paper [1], researchers analyzed SVM, logistic regression, decision tree, and random forest algorithms to assess which predicted Alzheimer onset most effectively on an OASIS dataset [1]. SVM was most accurate, while other models suffered from overfitting [1]. In paper [2], decision tree, random forest, SVM, gradient boosting, and voting classifiers were used on OASIS datasets to discover that dementia is more common in men than women, non-demented individuals have higher brain volume ratios than demented individuals, and there is a higher percentage of demented patients that are 70-80 years old than non-demented patients [2].

Our project's dataset is OASIS-1, which contains cross-sectional MRI data from both demented and non-demented individuals. Dimensions include gender, handedness, age, education, socioeconomic status, Mini-Mental State Examination score, Clinical Dementia Rating, Estimated Total Intracranial Volume, Normalized Whole Brain Volume, and Atlas Scaling Factor.

Link:
https://sites.wustl.edu/oasisbrains/files/2024/04/oasis_cross-sectional-5708aa0a98d82080.xlsx

Problem Definition

Alzheimer's affects nearly 24,000,000 people worldwide, with almost 500,000 new yearly cases in the United States alone [3]. Additionally, current human-centered diagnosis methods for early-stage Alzheimer's are unreliable [4]. Our project motivation is that ML algorithms could serve as a valuable technique to detect data indicative of "the subtle brain changes that occur in the preclinical stage of the disease" [4], to speed up diagnosis and ultimately help individuals better manage the symptoms and effects of Alzheimer's disease.

Methods

Pre-Processing

In our Project Proposal, we considered using the following three pre-processing methods:

1. Categorical Encoding (available in sklearn.preprocessing), which will be used to transform categorical data in the dataset into a numerical form compatible with our algorithms.

2. Standardization (available in sklearn.preprocessing), which will allow us to manipulate our data and fit it onto a standard normal distribution to improve future model accuracy.

3. Value Imputation (available in sklearn.preprocessing), which will be used to manipulate features which have incomplete information to render them usable in our models.

The pre-processing methods we have ended up implementing for the project are a categorical encoder built using sklearn.preprocessing's OneHotEncoder transformer, a standardizer built using sklearn.preprocessing's StandardScaler, and a value imputator to generate data points that were missing in the OASIS dataset.

Using sklearn's OneHotEncoder was necessary as certain dimensions in our dataset (namely the one corresponding to an individual's gender and handedness) are categorical and take on values such as M/F and L/R respectively. By using OneHotEncoder, we have converted the categorical values described above into binary equivalents. This makes the categorical data present in the OASIS dataset compatible with our models.

The StandardScaler in sklearn.preprocessing was used to implement Standardization and fit all non-binary and non-categorical data in the OASIS dataset onto a standard normal distribution. This was done to increase our data's consistency and make it easier for our models to interpret and make predictions in the future.

Additionally, we made use of MIDASpy's Variational Autoencoder to perform value imputation on the data. Performing value imputation was especially critical at this stage in the project as certain columns in the OASIS dataset have a lot of missing data, particularly for the columns referring to Educational Status (Educ), Socio-economic Status (SES), Mini-Mental State Evaluation (MMSE), and Clinical Dementia Rating (CDR). Performing value imputation allows us to substitute missing values in these categories based on statistical patterns drawn from the existing dataset as a whole. Since CDR was our target variable, we marked which values of CDR were imputed so that they did not impact our accuracy in testing.

We chose to use the MIDASpy Variational Autoencoder over more traditional value imputation in sklearn.preprocessing, such as the mean-mode imputer, following feedback from our TA, Richard, who suggested that the VAE is more robust. In general, Variational Autoencoders are able to "encode a continuous, probabilistic representation of that latent space, enabling not only accurately reconstructing exact original inputs, but also generation of new data samples that resemble the original input data using variational inference," [5] as opposed to traditional autoencoders, which only "encode a discrete, fixed representation of latent variables" [5]. Ultimately, a Variational Autoencoder provides a far more accurate way to impute large numbers of data points for the required columns than a more traditional method, ultimately leading to it being chosen.

Algorithms and Models Used

Beyond the pre-processing stage, the three ML algorithms/models identified in our proposal were:

1. SVM, by using sklearn.svm, we will leverage SVM's advantages in analyzing high-dimensional datasets while also validating the accuracy of the approach presented in paper [1].

2. KNN, by using sklearn.neighbors, we want to look at the subset of dimensions that relate to brain-imaging data and build a model to classify if an Alzheimer diagnosis can be made.

3. Random Forest, by using sklearn.ensemble, we want to implement our own random forest model to validate methods used in papers [1] and [2] while also determining its accuracy in predicting Alzheimer onset.

Random Forest

The first algorithm we have implemented is a Random Forest model using the sklearn.ensemble package's RandomForestClassifier method. The primary reasons behind why we chose to implement RandomForestClassifier are twofold. Firstly, it was one of the models used in papers [1] and [2] as one of the possible models to detect Alzheimer onset in patients. By implementing it in our project, we aimed to start working towards seeing if we could validate the performance of the model made in the aforementioned papers.

Secondly, Random Forest models have certain unique advantages which makes them particularly useful when dealing with datasets such as OASIS. First of all, The relationship between Clinical Dementia Rating (CDR) and predictor variables such as Estimated Total Intracranial Volume (ETIV) and Normalized Whole Brain Volume (NWBV) is likely to be complex and not effectively modellable using a more conventional linear model such as Linear Regression.

Random Forest typically serves as a better estimator than more conventional models such as the one discussed above.

Moreover, another defining feature of RandomForest is that it can be used in helping to determine which predictor variables are most significant and relevant for predicting CDR, which could yield useful information in determining what signs should be looked out for most frequently when diagnosing forms of dementia such as Alzheimer's disease. Ultimately, the combination of these factors has led us to prioritize the implementation of a RandomForest model using sklearn's RandomForestClassifier in our project.

SVM/SVC

The second algorithm implemented within the project was a SVM classification model (SVC) using the sklearn.svm package's SVC method. We chose to implement SVC for a variety of reasons. The primary one relates to the advantages SVM models have in multiple-class classification problems. SVM models work to "classify data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space" [6]. SVM models are particularly useful when we have multiple dimensions that are all relevant (in some capacity) towards the variable we are trying to predict. This is particularly applicable to our particular problem space, which has twelve different dimensions of data that can contribute to Alzheimer onset prediction in some capacity. SVM's usefulness when dealing with these high-dimensional datasets therefore makes it a logical choice to be used on the OASIS dataset, which ultimately motivated its selection.

Moreover, a variation of SVM was one of the models used in papers [1] and [2] as an alternative approach to detect Alzheimer onset in individuals. It ended up being the most accurate model, reportedly achieving " 69 correct predictions and only 6 wrong predictions with 92% accuracy" [1]. It also "possessed 91% of both test recall and test AUC, with no over-fitting or underfitting issues observed from the model" [1]. By implementing our own SVM implementation, we aim to validate the results obtained in our background literature and compare them to previous ML-based Alzheimer detection models alongside certain alternatives that the original papers did not consider, such as KNN, to ultimately determine if SVM remains the most accurate or if alternative predictors could prove more suitable for diagnosing Alzheimer's Disease.

KNN

The third algorithm implemented is a K-Nearest Neighbors (KNN) model using KNeighborsClassifier from the sklearn.neighbors package. Unlike the other two models we implemented, KNN provides a more simple approach to classification by leveraging distances between k-nearest data points. Because of its simplicity, we realized that it could be a good baseline as a comparison for the other complex models. KNN is also effective at handling

high-dimensional datasets, such as our OASIS dataset, thus we chose this in addition to SVM. To optimize our model, we used hyperparameter tuning GridSearchCV to identify the best combination of k neighbors, distance metrics, and weighting strategy. Since KNN is sensitive to differences in feature scales, we used StandardScaler to standardize features to ensure a fair comparison.

Our evaluation of KNN included metrics such as accuracy, precision, recall, F1 score, and a confusion matrix for visualizing classification performance. We also included a stratified 5-fold cross-validation to assess the robustness of the model, which provides additional insight into its performance stability across different data splits.

Ultimately, we included KNN as a way to meaningfully compare Random Forest and SVM. It is a benchmark model to validate if a simpler model can achieve high accuracies on an Alzheimer's dataset compared to parameterized complex models. Overall, KNN complements the other models in our study, helping us evaluate different approaches to diagnosing Alzheimer's disease.

Results And Discussion

Random Forest

The first algorithm implemented in our project was the Random Forest model using the OASIS-1 dataset. The preprocessing pipeline included OneHot encoding for categorical data, standardization of numerical features, and value imputation using a Variational Autoencoder (VAE) from MIDASpy. These steps helped address missing values and ensured compatibility with the machine learning algorithms.

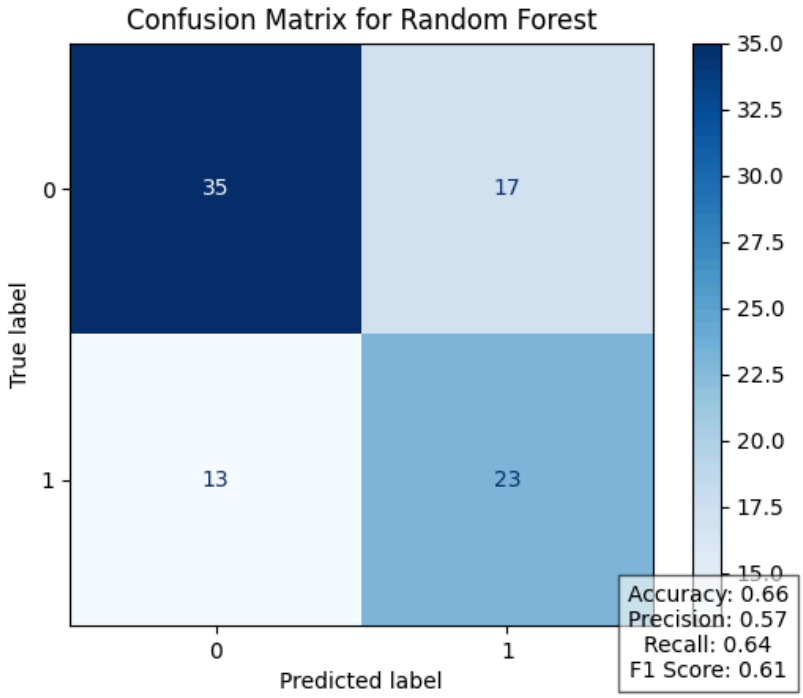
Quantitative Metrics

The Random Forest model achieved the following performance metrics:

- Accuracy: 66%
- Precision: 57%
- Recall: 67%
- F1 Score: 61%

The model's accuracy is 66%, indicating moderate predictive capability. The recall of 64% suggests that the model can correctly identify 64% of the individuals who have Alzheimer's onset, which is critical given the goal of minimizing missed cases. However, the relatively low precision of 57% indicates a higher rate of false positives, where the model incorrectly predicts Alzheimer's onset.

Confusion Matrix Analysis

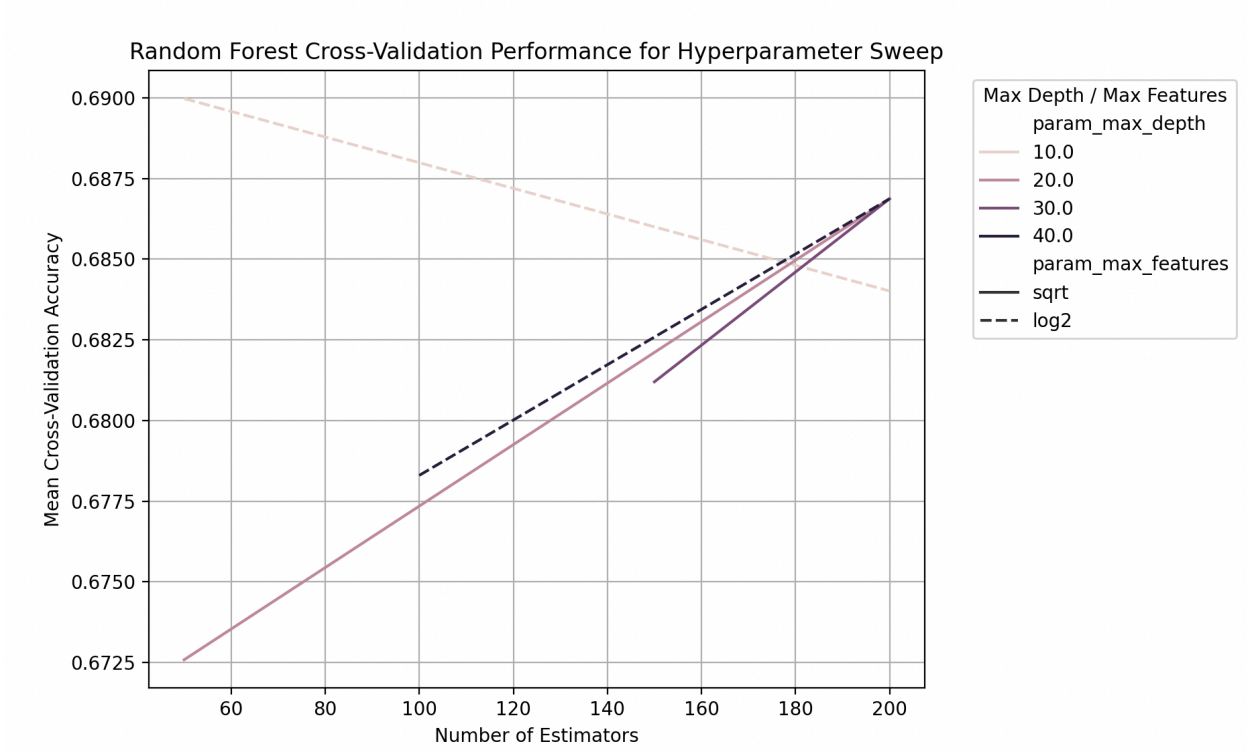


The confusion matrix provides a detailed breakdown of the model's predictions:

	Predicted No Onset	Predicted Onset
Actual No Onset	35 (True Negative)	17 (False Positive)
Actual Onset	13 (False Negative)	23 (True Positive)

The model misclassified 13 actual onset cases as 'No onset', which is concerning given the importance of identifying early Alzheimer's cases. Additionally, the model had 17 false positives, predicting onset where there was none. This imbalance may indicate issues with overfitting or a need for a better handling of class distribution in the dataset.

Hyperparameter Sweep Analysis



The x-axis represents the number of estimators in a random forest, so the number of trees. Our testing used a range from 50 to 200 estimators. The y-axis represents the average accuracy cross-validation for each configuration tested. This shows how well each hyperparameter generalizes across subsets of data.

The legend indicates different max depths and features settings used for the trees. Lighter colors represent lower depths (10), while darker colors represent higher depths (40). The solid lines represent a square root being used to determine the number of splits, while dash lines represent a base-2 logarithm to determine features.

Generally, we see a positive upwards trend for cross-validation accuracy. as the number of estimators increases. This means that adding more trees will typically result in an improved accuracy. Taking a look at the max depth parameters, we notice that deeper trees resulted in a similar positive upwards trend for accuracy. In fact, the shallowest depth, 10, resulted in a decrease in accuracy as the number of estimators increased. Analyzing the features method, we see that using a square root is generally better than using base-2 logarithmic since the former has a greater slope. We can reason that this is due to a larger subset of features being considered, allowing more patterns to be recognized in the data. The highest accuracy values are achieved when we have higher values of estimators, high max values, and a square root to determine max features.

Discussion

The hyperparameter tuning analysis provided valuable insights into the model's behavior and helped identify configurations that improve its performance. However, despite these optimizations, the model still exhibits a relatively low precision score of 57%, indicating a higher rate of false positives. This trade-off affects the model's reliability in real-world scenarios where incorrectly diagnosing Alzheimer's could lead to undue stress for patients.

Several factors may be influencing these results:

1. Class Imbalance: The dataset may have a higher proportion of non-onset cases, potentially biasing the model towards predicting 'No onset'
2. Feature Relevance: Certain features may not be contributing effectively to the prediction, leading to noise in the model's decisions.
3. Overfitting: The random Forest model may be capturing noise in the training data rather than general patterns, leading to decreased performance on the test set.
4. Dataset Limitation and Potential Bias: While we have implemented robust imputation techniques using a Variational Autoencoder (VAE), it is important to note that the observed CDR values in our dataset are predominantly for individuals aged 60 and above. As a result, younger individuals lack real CDR data, and their values were imputed. This introduces a potential source of bias, as the imputed values may not accurately represent the true distribution of dementia ratings for younger individuals. We acknowledge that this limitation could affect our model's predictions, leading to possible overfitting or incorrect classifications for younger age groups.

SVM/SVC

The second algorithm implemented within our project was a SVC classifier, which is a variation of a Scalable Vector Machine for multi-class classification problems implemented using ScikitLearn's SVC package on the OASIS-1 dataset. Alongside this algorithm, we implemented hyperparameter tuning using a technique known as Grid Search, which allowed us to use a variation of SVC with the combination of parameters that would yield the highest possible accuracy. The results of the model are presented below.

Quantitative Metrics

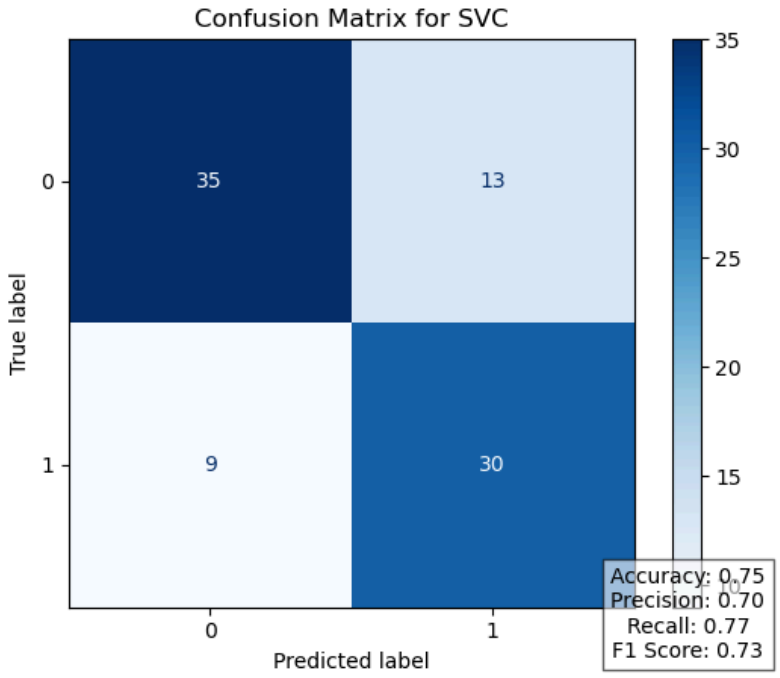
Our SVC model achieved the following scores based on the project's primary performance metrics:

- Accuracy: 75%
- Precision: 70%
- Recall: 77%
- F1 Score: 73%

Following from the analysis conducted for our Random Forest algorithm, the SVC model's accuracy is 75%, which represents an increase over the Random Forest model's accuracy of 66%. The recall of 77% indicates that our SVC model correctly identifies 77% of the individuals who have Alzheimer's onset. This is again an improvement over the Random Forest algorithm, and tells us that the SVC model is a better option when it comes to reducing the number of instances where someone suffering from Alzheimer's disease does not get diagnosed, which is arguably the most serious wrong prediction the model can make. The precision of 70% was also an improvement based on the previous model, meaning that the SVC model is less prone to misdiagnose a case as a false positive (predicting that someone has Alzheimer's when they do not) than the previous one.

Overall, while the SVC model generally performed better than our original Random Forest, the initial values for accuracy and recall are still below the 80% and 85% thresholds identified at the beginning of the project. A visualization of the model's results using a Confusion Matrix is presented below:

Confusion Matrix



The confusion matrix presented above shows a more visual representation of the SVM model's overall performance:

	Predicted No Onset	Predicted Onset

Actual No Onset	35 (True Negative)	13 (False Positive)
Actual Onset	9 (False Negative)	30 (True Positive)

The overall number of False Negatives and False Positives the model misclassified was 9 and 13 respectively. This is, once again, an improvement over the previous model. It nonetheless could be considered somewhat unsatisfactory; as Alzheimer's is a disease which brings about life-changing impacts and consequences, the number of False Positive/False Negative diagnosis results should be minimized as much as possible, leaving room for further improvement.

Discussion

Before beginning, it is important to consider that the parameters of the SVC model fitted onto the data were determined using a hyperparameter tuning method known as GridSearchCV which is implemented within Scikit-Learn. GridSearch allowed us to specify different kernels, regularization coefficients, gamma values, polynomial degrees, and bias coefficients. It would then run an algorithm to determine the combination of parameters that yielded the highest possible accuracy. The parameters that the GridSearch method was using to fit the best possible SVC model were:

```
parameters = {
    "kernel" : ['linear', 'rbf', 'poly', 'sigmoid'],
    "C" : [0.1, 1, 10, 20, 50, 100],
    "gamma" : ['scale', 'auto'],
    "degree" : [2, 3, 4],
    'coef0' : [0, 0.1, 0.5, 1, 10]
}
```

The GridSearch algorithm we ended up running demonstrated that the best possible SVC model used a polynomial kernel with degree 2 (so a quadratic kernel was used), and a regularization coefficient of 10. The gamma value was irrelevant in this particular instance given that a polynomial kernel ended up being applied to the dataset. Beyond this, we also performed K-Fold Cross Validation using 5 folds to determine the model's accuracy on unseen data and see the risk for overfitting. The accuracy values associated with each fold were relatively consistent, being in a range between 70% and 78%, which provides evidence that the SVC model likely was not notably overfitting data values.

Overall, the SVC model yields a balanced set of results when it comes to evaluating its accuracy, precision, and recall metrics. The values of 75% for accuracy, 70% for precision, and 77% for recall are still nonetheless somewhat low given the nature of the problem space and how critical

it is to diagnose Alzheimer's Disease correctly in many different situations and contexts, which demonstrates there could be room for possible improvements to be made.

One such way to improve the performance of our SVM classifier could be by closely evaluating the importance of certain data categories in predicting Alzheimer's. Data such as the handedness of a participant or their gender is included in the original dataset, but it may be the case that these columns are not as important in detecting Alzheimer Onset as ones such as Estimated Total Intracranial Volume (ETIV) and Normalized Whole Brain Volume (NWBV). The importance of each column could therefore be considered and modelled to allow the model to consider it and improve the overall accuracy, precision, and recall metrics.

Additionally, another acknowledgement that must be made relates to the nature of our dataset and the lack of Clinical Dementia Ratings (CDR) associated with younger individuals present in it. As has been mentioned previously, younger individuals in the OASIS dataset do not have CDR scores. We had to impute these values using a Variational Autoencoder, which could introduce bias within the dataset and lead the model to be more prone to making incorrect predictions for individuals in younger age groups.

KNN

The third algorithm implemented is K-Nearest Neighbors. KNN was implemented using the KNeighborsClassifier from the sklearn.neighbors package to serve as a baseline for our Alzheimer's onset prediction. By calculating the distances between k-nearest data points, the model provides a straightforward approach to classifying whether or not an individual is at risk of developing Alzheimer's based on the OASIS dataset.

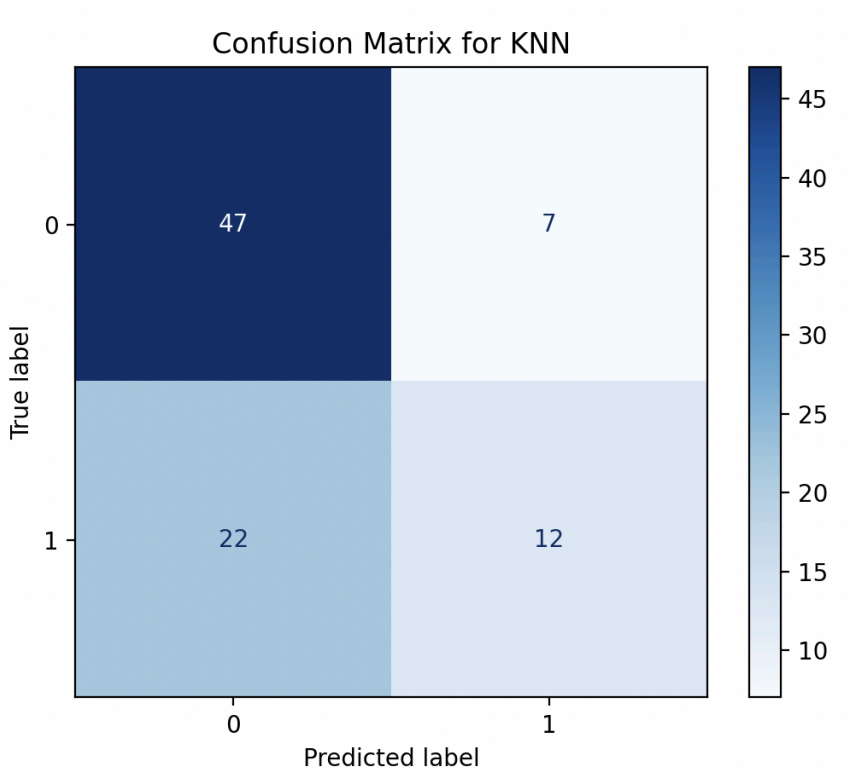
Quantitative Metrics

The KNN model achieved the following performance metrics:

- Accuracy: 67%
- Precision: 63%
- Recall: 35%
- F1 Score: 45%

While KNN demonstrated moderate accuracy and precision, its low recall suggests that the model struggled to correctly identify individuals with Alzheimer's onset. This imbalance is also reflected in the F1 score. Compared to performance metrics from the SVM model, this indicates a limited effectiveness in balancing precision and recall.

Confusion Matrix



The confusion matrix presented above shows a more visual representation of the KNN model's overall performance:

	Predicted No Onset	Predicted Onset
Actual No Onset	47 (True Negative)	7 (False Positive)
Actual Onset	22 (False Negative)	12 (True Positive)

The confusion matrix for the KNN model shows that it correctly classified 47 true negatives and 12 true positives but misclassified 22 false negatives and 7 false positives. While the model performed moderately well in identifying non-onset cases, the high number of false negatives (22) highlights its difficulty in correctly identifying individuals with Alzheimer's onset, reflected in the low recall score of 35%. This imbalance is critical, as false negatives pose a significant risk in medical diagnosis. Addressing this issue through techniques like class weighting or data augmentation could improve the model's recall and overall performance as a point for further improvement.

Discussion

The KNN model's performance metrics reflect its simplicity and limitations in addressing the complexity of Alzheimer's onset prediction. While it achieved moderate accuracy (67%) and precision (63%), the low recall score (35%) highlights the model's difficulty in correctly

identifying individuals with Alzheimer's onset. This issue is particularly concerning in the context of medical diagnosis, where false negatives can lead to missed cases and delayed treatment. The F1 score of 45% further underscores this imbalance, as the model struggles to balance precision and recall effectively.

Several factors may have contributed to KNN's limited performance. As a distance-based model, KNN relies heavily on the feature space's geometry, making it sensitive to class imbalance and irrelevant features. High-dimensional datasets like OASIS can dilute the effectiveness of distance metrics, especially without robust feature selection. Moreover, the relatively low recall may indicate that the model struggles to identify subtle patterns in the data, which could be addressed through more advanced preprocessing techniques or the incorporation of domain-specific features.

Hyperparameter tuning via GridSearchCV allowed us to explore different combinations of k, distance metrics, and weighting strategies, but the results suggest that KNN may not fully capture the complexity of relationships between features and Alzheimer's onset. Additionally, the confusion matrix reveals that while KNN performs reasonably well in identifying non-onset cases (47 true negatives), it struggles significantly with onset cases, with 22 false negatives and only 12 true positives.

To improve KNN's performance, future steps could include addressing class imbalance through techniques like SMOTE, conducting thorough feature selection to focus on the most relevant predictors, and experimenting with advanced distance metrics or kernelized approaches. While KNN serves as a valuable baseline for comparison, its results highlight the need for more sophisticated models like SVM and Random Forest to address the challenges of Alzheimer's onset prediction effectively.

Final Model Comparison

The three models implemented, Random Forest, SVM, and KNN, represent a diverse range of machine learning approaches, each with unique strengths and limitations. By comparing their performance metrics, we gained insights into how well each model addresses the challenges of predicting Alzheimer's onset using the OASIS-1 dataset.

	Random Forest	SVM/SVC	KNN
Accuracy	66%	75%	67%
Precision	57%	70%	63%
Recall	67%	77%	35%

	Random Forest	SVM/SVC	KNN
Accuracy	66%	75%	67%
F1 Score	61%	73%	45%

Accuracy

SVM achieved the highest accuracy (75%) among the models, reflecting its ability to handle high-dimensional data and find optimal decision boundaries. KNN and Random Forest followed with 67% and 66% accuracy, respectively.

Precision

Precision measures the proportion of true positives among all predicted positives, and SVM outperformed the other models with a precision of 70%. This suggests that SVM is less prone to false positives compared to Random Forest (57%) and KNN (63%). KNN performed moderately well in this regard, reflecting its ability to make relatively accurate predictions for positive cases.

Recall

Recall is critical in Alzheimer's prediction because failing to identify individuals at risk can have significant consequences. SVM excelled in this metric with a recall of 77%, followed by Random Forest at 67%. KNN struggled with a recall of 35%, indicating its difficulty in identifying true positive cases.

F1 Score

The F1 score balances precision and recall, offering a single metric to compare models. SVM achieved the highest F1 score (73%), reinforcing its status as the best-performing model overall. Random Forest achieved a moderate F1 score of 61%, while KNN's low recall contributed to its comparatively poor F1 score of 45%.

Discussion

SVM emerged as the most robust model, offering the best balance across accuracy, precision, recall, and F1 score. Its ability to handle high-dimensional data and identify subtle patterns in the OASIS dataset made it the most effective algorithm for predicting Alzheimer's onset. Random Forest, while less accurate than SVM, demonstrated strong recall, making it a viable option for minimizing missed cases. However, its lower precision indicates a tendency to generate false positives, which could cause undue stress in a real-world diagnostic setting. KNN, despite its simplicity and moderate precision, underperformed due to its low recall, making it quite unreliable for this problem.

Next Steps

Based on the hyperparameter analysis and current results, these steps would help improve our final model:

1. Handle Class Imbalance: Experimenting with techniques like SMOTE or adjusting class weights to mitigate the impact of class imbalance. This will improve recall and reduce false negatives, especially for KNN and Random Forest.
2. Feature Selection: Conducting a thorough feature importance analysis and potentially remove features that contribute little to the model's predictive power.
3. Ensemble Techniques: Exploring ensemble methods such as stacking and voting classifiers to combine the strengths of Random Forest, SVM, and KNN, potentially reducing false positives and improving overall accuracy.
4. Look to expand the dataset: Seeking access to other publicly available datasets on Alzheimer's to increase sample size and improve our models.

References

[1] M. Bari Antor et al., "A comparative analysis of machine learning algorithms to predict alzheimer's disease," Journal of Healthcare Engineering, vol. 2021, pp. 1–12, Jul. 2021. doi:10.1155/2021/9917919

[2] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, "Early-stage alzheimer's disease prediction using machine learning models," Frontiers in Public Health, vol. 10, Mar. 2022. doi:10.3389/fpubh.2022.853294

[3] "Alzheimer's Facts & Statistics," Alzheimer's San Diego, https://www.alzsd.org/resources/facts-stats/ (accessed Oct. 3, 2024).

[4] R. Mistur, L. Mosconi, D. S. Santi, M. Guzman, Y. Li, W. Tsui, and M. J. de Leon, "Current challenges for the early detection of Alzheimer's disease: Brain imaging and CSF studies," *J. Clin. Neurol.*, vol. 5, no. 4, pp. 153-166, Dec. 2009. doi: 10.3988/jcn.2009.5.4.153.

[5] D. Bergmann and C. Stryker, "What is a Variational Autoencoder?," IBM, https://www.ibm.com/think/topics/variational-autoencoder (accessed Nov. 11, 2024).

[6] IBM, What are SVMs?, IBM, https://www.ibm.com/topics/support-vector-machine (accessed Dec. 2, 2024).

Gantt Chart

Link to Gantt Chart: Project_Gantt_Chart.xlsx

Final Milestone Contributions

Member	Contribution
Tommaso	Final Coding: SVM Implementation, Final Report: SVM Method, SVM Discussion, Slides: Slide Creation
Ion	Final Report: KNN Method, KNN Discussion, Final Model Comparison, Next Steps
Kathryn	Slides, Final Recording
Sriteja	Random Forest Implementation, Slides, Final Recording
Jude	Slides, Final Report, Final recording
Nikhil	Final Coding: KNN Implementation

GitHub Repo Link

https://github.gatech.edu/tadani3/fall-24-ml-project