MULTIOMICS BASED CLASSIFICATION OF PREDIABETIC PATIENTS USING SUPERVISED VAE DRIVEN DATA COMPRESSION

Advait Padhye[a], Aparna Chauhan[a], Advait Parmar[a], Raghavendra Lakshminarayanan[a], Kshitij Jadhav[a]

[a]*KCDH, IIT Bombay, Powai, Mumbai, 400076, Maharashtra, India*

Abstract

Identification of diabetic tendencies in an individual is dependent on multi-factorial omic modalities among other clinical characteristics. Distinction of prediabetes patients has additional difficulty because even though characterized by higher than normal blood glucose levels, other features are difficult to delineate from healthy state. However, prediabetes does indicate an increased risk of developing diabetes and thus is important to classify. Different omic data contribute to the disease pathology and multiple features can be the indicator of the patient's fate. But, multiomics data generation is an expensive and complex task, leading to difficulty in profiling large cohorts, further causing data availability hurdle. Additionally, omic profiling technologies are subject to bias in acquiring data, sensitivity of the instrument and the high complexity in biological samples. Hence, missing features are inevitable attribute of each omic block, while an entire block of data might be absent occasionally due to various technical and manual constraints. The aforementioned aspects add to the impediment in accurate classification of samples into healthy and prediabetes, thus hindering the ability for early intervention and prevention of people with T2 diabetes predisposition in retroverting to healthy state. This paper explores a novel segmented missing value imputation strategy for multiomics datasets to improve patient categorization. Subsequently, an in house developed supervised clustering module with modified variational auto encoder is used for prediction of prediabetic individuals. The model is trained on proteomic, metabolomic, gut microbiome and clinical data of 106 individuals from the iHMP cohort. The use of segmented KNN imputation on the dataset gave an improved prediction accuracy of 0.947 with supervised VAE in comparison to 0.914 accuracy without imputation. Also, our model showed an improved accuracy with and without imputation when compared to traditional VAE and multilayer perceptron models. The highest performance of DC SAE among the other state-of-the art approaches indicate the effectiveness of the model for binary classification from high dimensional data. In conclusion, a modified kNN imputation followed with DC SAE based binary classification demonstrated feasibility for accurate prediction of prediabetes from a small and sparse data while also reducing the dimension of predictive features using revamped VAE, further reducing the risks involved in patient data protection.

Keywords: prediabetes, multiomics, deep learning, auto encoder

1. Introduction

The burden of diabetes is on the rise worldwide, particularly in developing economies such as India, primarily due to the increasing occurrence of overweight/obesity and unhealthy lifestyles. According to 2019 estimates, India had 77 million individuals living with diabetes, and this number is projected to soar to over 134 million by 2045. Alarmingly, approximately 57 percent of these individuals remain undiagnosed (Pradeepa and Mohan, 2021). The global prevalence of diabetes has also been rising steadily, and it has now reached epidemic proportions, posing a significant burden on healthcare systems and economies. Uncontrolled diabetes can lead to a myriad of complications, including heart disease, stroke, kidney failure, blindness, and lower limb amputations, all of which have profound implications for both the individual's quality of life and healthcare costs.

Prediabetes is a condition characterized by higher than normal blood glucose levels but not yet meeting the criteria for a diabetes diagnosis. Detecting prediabetes patients is of paramount importance in combating the increasing number of diabetic cases, as it represents a critical opportunity to intervene and prevent the progression to full-blown diabetes. This stage serves as a crucial window of opportunity for early intervention, lifestyle modifications, and medical management to halt the onset of diabetes. Early detection of prediabetic patients will also allow healthcare professionals to implement targeted treatments that can effectively lower blood glucose levels, reduce the risk of developing diabetes, and prevent associated complications. Lifestyle modifications, such as dietary changes and increased physical activity, are often the first-line approach, which can have a substantial impact on improving insulin sensitivity and glucose metabolism.

Moreover, identifying prediabetic individuals is not just about averting diabetes. It also serves as a means to address the root causes of this condition, including obesity, poor dietary habits, sedentary lifestyles, and genetic predisposition. By getting involved at this early stage, healthcare providers can work with patients to address these underlying factors and promote healthier behaviours, which can have far-reaching benefits for their overall well-being. From a public health perspective, the prediction of prediabetes is crucial for resource allocation and preventive strategies. It enables healthcare systems to target high-risk populations, implement educational campaigns, and allocate resources efficiently to manage and mitigate the growing diabetes epidemic.

Despite the ongoing research, it is not only difficult to predict individuals with a tendency to develop diabetes, but more so to estimate the fate of a patient. A variety of Machine learning (ML) and Deep learning (DL) algorithms have emerged as powerful tools in predicting disease conditions (Ma et al. (2020), Reel et al. (2022), Hu et al. (2022), Zhuang et al. (2023)) using omics data, however such data is often subject to various problems ranging from data non-availability and missing values to heterogeneity across and within groups. This leads to state-of-art approaches performing poorly especially in case of high dimensional datasets. The challenge of capturing intricate non-linear patterns in high-dimensional data has been solved with emergence of robust methodology provided by Deep Learning models (LeCun et al., 2015). Within the realm of contemporary deep learning techniques, the Variational Autoencoder (VAE) (Pinheiro Cinelli et al., 2021) has risen as a promising solution for embedding omics data into a lower-dimensional latent space. When combined with a downstream classification network, the VAE-based models have demonstrated good performance in the categorization of biological samples and interactions, surpassing alternative machine learning and deep learning approaches (Azarkhalili et al. (2019) , Zhang et al. (2019), Hira et al. (2021), Zhang et al. (2021)). A VAE being an unsupervised algorithm is currently suited mostly for generative modelling, however, recently a modified approach was developed for binary classification [put DC SAE reference] using a supervised VAE module. The approach combined with a novel strategy of segmented missing value imputation using kNN led to a significant improvement in segregation among healthy controls and prediabetic patients. Furthermore, better classification could be achieved from imputed dataset in comparison to a non-imputed dataset for all the tested DL algorithms signifying the importance of Missing Value Imputation (MVI) in multimodal data.

2. Materials and Methods

2.1. Dataset Selection

Based on the dynamic and importance of features in disease pathology the presence of proteomics, metabolomics, gut microbiome, and clinical profiles was essential. The following study - http://med.stanford.edu/ipop.html hosted on iPOP site of Stanford which is also a part of the iHMP effort was selected because it met the criteria of including the required profiles and sufficient metadata was also provided enabling integration of the different features and longitudinal samples. The dataset includes multiomics profile and clinical characteristics for 106 samples of which 87 are prediabetic and remaining are healthy controls. The proteomics and metabolomics data were collected from serum samples. The gut microbiome abundance was profiled from the faecal samples. The data was gathered for about 3 years with each patient having varying number of visits some of which also included confounding factors such as infection stage and antibiotic intake.

2.2. Integration

As the dataset consisted of longitudinal data from different omic experiments, we started with integrating the healthy visit of each patient. All the visit for each sample was taken and filtered for all confounding effects, after which a trimmed mean was taken of every feature for patients having higher than 10 healthy visits, otherwise a simple average was calculated. For inclusion of categorical clinical characteristics, they were converted to binary or continuous variables based on implication of the clinical variable measured on prediabetic person.

2.3. Dataset Problems

Inspite of the huge potential that multi-omics data possesses in distinguishing between pre-diabetic and healthy patients, often times, these datasets are plagued with several issues. Most of these issues stem from the fact that there is an inherent bias during data acquisition process coupled with the sensitive nature of the instruments used in the omic profiling technologies. Such factors result in various missing features in each omic block and occasionally an entire block of omics data might be missing in the dataset. Traditional imputation techniques such as replacing with 0 or mean value of the feature result in poor performance of the downstream classification tasks. To combat this problem, we propose a novel **Segmented Imputation** technique that employs k-Nearest Neighbour algorithm to impute missing values in the omics dataset.

In addition to the missing value problem, disproportionate sample numbers between classes is a common observation in omics datasets. The dataset that we used for our analysis, consisted of 85 prediabtic samples and only 19 healthy controls. Such an imbalance between data samples introduces a biasness in the classification models which results in their poor performance on unknown data samples. To mitigate this issue of data imbalance, we employed oversampling techniques involving SMOTE (Synthetic Minority Over-sampling Technique).

2.4. Segmented kNN imputation

Multi-Omics datasets usually consist of multiple blocks containing different omics data. The dataset that we used in our analysis, consisted of 4 blocks containing metabolomics, proteomics, gut microbiome and clinical data. Data in each of these blocks have distinct patterns and underlying distributions. Due to this, performing imputation using data from blocks other than the one where the feature is missing, will result in an imputed value which is inconsistent with the underlying distribution of that particular block. This can be overcome by initially performing imputation within blocks and then across the blocks. This formed the basis of our proposed segmented imputation technique.

For performing imputation, we employ k-Nearest Neighbour algorithm. For a data sample containing a missing feature, we first find its nearest neighbours where that particular feature is not missing and then impute the missing value with the average of feature values from the neighbours.

2.4.1. *Preprocessing*

Preprocessing of the data was necessary to combat the imputation problem when high percentage of features were missing for a sample.So,a pruning percentage was defined, $p \in [0, 100]$ based on which we removed certain data samples and features. For our analysis, we set $p = 90$. Based on the value of p, we pruned out the data samples for which more than $p\%$ features were missing. Similarly, we removed those features which were missing in more than $p\%$ data samples. This filtering was done to ensure that data samples and features in our classification model which didn't possess much information to assist in the classification task were excluded from the input.

2.4.2. *Within Block Imputation*

The original dataset was first divided into blocks each containing the respective omics data. In our analysis, since we are working with metabolomics, proteomics, gut microbiome and clinical data, the dataset was segregated into 4 blocks. The kNN imputation was be applied separately to each of these blocks. In this step, we only considered features belonging to one particular omic block while performing imputation.

We defined another quantity called block limit, $b \in [0, 1]$. Based on this quantity, we decided whether imputation could be performed using the data present only within this block or we would need to consult data from other blocks as well. In our analysis, we set $b = 0.5$ meaning that for a data sample, if more than half of its features were missing then imputation could no be completed using data only from one omic block. Such data samples were kept as such and forwarded to the next segment for imputation.

We first identified the set of data samples which contained missing feature values and based on b only consider those data samples which could be imputed from the features within the same block. During this entire procedure, we intended to impute only one missing feature at a time. We denoted $\mathcal{N}$ as the set of neighbouring samples that were to be used for imputation of a missing feature.

After fixing the data sample and the missing feature to be imputed, we looked for other data samples (which we called $\mathcal{D}$) where that feature was not missing. Here, we defined another quantity called row threshold $r \in [0, 1]$ which was used to determine whether the set of k nearest neighbours (denoted by $\mathcal{D}_k$) would be computed within $\mathcal{D}$ or the entire set would be utilized. In the first case, $\mathcal{N} = \mathcal{D}_k$ and in the second case $\mathcal{N} = \mathcal{D}$. The value of r was set to 0.2 which implied that if $\mathcal{D}$ consisted of more than one-fifth of the total number of data samples then the k nearest neighbours were computed instead of using the entire set.

There was a small subtlety while implementing KNN algorithm in this setup. Few members of the set $\mathcal{D}$ themselves had some features missing other than the one currently being imputed.To implement KNN algorithm in such cases, these missing features needed to be imputed. In this case, we simply imputed these missing features with the mean value of that feature across all data samples.

Finally after determining $\mathcal{N}$, we imputed the missing feature with the mean value of that feature within the set $\mathcal{N}$. This procedure was repeated for each missing feature in each data sample within each omics block.

2.4.3. *Across Block Imputation*

As mentioned in the previous section, there were some data samples whose number of missing features exceeded the block limit b. These remaining data samples were imputed in this step where features from all the omics blocks were considered. Like previous step, we imputed only one missing feature at a time.

As described in the procedure of previous segment this step also started with determination of the set $\mathcal{D}$ where the features required for imputation were not missing. Based on the value of row threshold r, we deduced the set $\mathcal{N}$ that would be used for imputation. The nuance mentioned in the previous step was also applicable in the second segment of imputation, hence the missing features other than the current feature were imputed with the mean value of that feature across all data samples. Finally, we imputed the missing feature with the mean value of that feature within set $\mathcal{N}$. This procedure was repeated for each missing feature in each remaining data sample.

2.5. *Oversampling using SMOTE*

Multiomics dataset are susceptible to imbalance due to several reasons such as disease patients' higher frequency of visiting for clinical tests in comparison to healthy patients. Such disparity results in poor discriminative ability of the classification models. To overcome this problem, we employed oversampling techniques.

Oversampling techniques involve increasing the number of minority class examples in the training set. A few examples of these techniques include Resampling, Random Oversampling, etc. SMOTE (Synthetic Minority Over-sampling Technique) was utilized for oversampling in this study.

In SMOTE, the minority class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors where k is a hyperparameter.

2.6. *Deep learning models*

Our main objective was to perform binary classification using the multi-omics dataset. The higher dimensionality and non-linearity present in multi-omics dataset necessitated the use of deep learning approaches for performing the classification tasks. In particular, we used **Multilayer Perceptron** and **Variational Autoencoders** for performing binary classification.

Multilayer Perceptrons (MLPs) represent the fundamental archetype of neural networks, finding application in regression and classification tasks. In our case, we employed a single hidden layer MLP classifier model as a foundational benchmark to assess the efficacy of the Variational Autoencoder (VAE) and Discrete Classifier Supervised Autoencoder models.

Variational Autoencoder is usually used for its generative aspects since it models a latent distribution from which we can sample more data. However, in this case, we used the latent space reprsentations of the input data obtained from the Variational Autoencoder for training a classifier model - Support Vector Machine (SVM).

2.7. *Discrete Classifier Supervised Autoencoder*

In the paper [DC-SAE paper], our lab had proposed a variant of Variational Autoencoder called **Discrete Classifier Supervised Autoencoder** (DC-SAE) which models two latent distributions, instead of one, each corresponding to one of the two classes. This was achieved by introducing a new loss term and switching to supervised learning setting which helped in generating better discriminative representations for classification.

For our classification task, we trained DC-SAE on the training dataset and obtain the latent space representations of this training dataset. Using these representation, we trained a Support Vector Machine (SVM). This SVM performed binary classification with the aim to segregate healthy and prediabetic samples.

3. Results

3.1. *Data Description*

A total of 829 measurements with healthy baseline profiles with no major confounding variables were obtained for the 106 samples with the healthy visit number varying between 1-56. After integration total number of samples and features, all the omics block of features combined gave a single high dimensional matrix consisting of 1173 features and 104 samples (2 samples removed in preprocessing). A total of 12040 features were missing in the entire sample space, i.e., approximately 113 features were missing on an average with missing features ranging between 0-1173 per sample. The dataset comprised 19 individuals categorized in healthy controls and remaining 87 had HbA1C levels in prediabetic/diabetic range hence were classified as prediabetic patients.

3.2. *kNN imputation results*

In preprocessing, two samples (ZOZOW1T and ZO94RDZ) from the prediabetic category were excluded due to all features being missing. x features were removed from the input data as they were missing for more than 90% samples. z number of features were imputed in within block segment, while q features were imputed in the across block segment of the imputation method. The effect of imputation on the ability of the multiomics data to classify the samples was studied by training and testing the deep learning models on original and imputed data.

3.3. *Balance in healthy and prediabetic numbers achieved with oversampling*

The presence of only 19 healthy controls in comparison to 85 prediabetes samples made the dataset predisposed to bias in classification model. To combat this problem, oversampling was performed using SMOTE for healthy controls to achieve a balanced dataset of z samples consisting of x healthy and y prediabetic individuals. Deep learning models were tested for the original and oversampled data to signify the importance of data augmentation in such studies.

3.4. *Deep Learning Models' results for original data*

3.5. *Deep Learning Models' comparison for imputed data*

4. Discussion

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

5. Summary and conclusions

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Acknowledgements

Thanks to ...

References

Azarkhalili, B., Saberi, A., Chitsaz, H., Sharifi-Zarchi, A., 2019. DeePathology: Deep Multi-Task Learning for Inferring Molecular Pathology from Cancer Transcriptome. Scientific Reports 9, 16526. doi:10.1038/s41598-019-52937-5.

Hira, M.T., Razzaque, M.A., Angione, C., Scrivens, J., Sawan, S., Sarker, M., 2021. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Scientific Reports 11, 6265. doi:10.1038/s41598-021-85285-4.

Hu, Y., Zhao, L., Li, Z., Dong, X., Xu, T., Zhao, Y., 2022. Classifying the multi-omics data of gastric cancer using a deep feature selection method. Expert Systems with Applications 200, 116813. doi:10.1016/j.eswa.2022.116813.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. doi:10.1038/nature14539.

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., Song, F., 2020. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in Biology and Medicine 121, 103761. doi:https://doi.org/10.1016/j.compbiomed.2020.103761.

Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E.A., Lima Netto, S., 2021. Variational Autoencoder, in: Variational Methods for Machine Learning with Applications to Deep Networks. Springer International Publishing, Cham, pp. 111–149. doi:10.1007/978-3-030-70679-1_5.

Pradeepa, R., Mohan, V., 2021. Epidemiology of type 2 diabetes in India. Indian Journal of Ophthalmology 69, 2932. doi:10.4103/ijo.IJO_1627_21.

Reel, P.S., Reel, S., van Kralingen, J.C., Langton, K., Lang, K., Erlic, Z., Larsen, C.K., Amar, L., Pamporaki, C., Mulatero, P., Blanchard, A., Kabat, M., Robertson, S., MacKenzie, S.M., Taylor, A.E., Peitzsch, M., Ceccato, F., Scaroni, C., Reincke, M., Kroiss, M., Dennedy, M.C., Pecori, A., Monticone, S., Deinum, J., Rossi, G.P., Lenzini, L., McClure, J.D., Nind, T., Riddell, A., Stell, A., Cole, C., Sudano, I., Prehn, C., Adamski, J., Gimenez-Roqueplo, A.P., Assie, G., Arlt, W., Beuschlein, F., Eisenhofer, G., Davies, E., Zennaro, M.C., Jefferson, E., 2022. Machine learning for classification of hypertension subtypes using multi-omics: A multi-centre, retrospective, data-driven study. eBioMedicine 84, 104276. doi:10.1016/j.ebiom.2022.104276.

Zhang, X., Xing, Y., Sun, K., Guo, Y., 2021. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. Cancers 13, 3047. doi:10.3390/cancers13123047.

Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., Guo, Y., 2019. Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 765–769. doi:10.1109/BIBM47256.2019.8983228.

Zhuang, Y., Xing, F., Ghosh, D., Hobbs, B.D., Hersh, C.P., Banaei-Kashani, F., Bowler, R.P., Kechris, K., 2023. Deep learning on graphs for multi-omics classification of COPD. PLOS ONE 18, e0284563. doi:10.1371/journal.pone.0284563.