Schizophrenia and Machine Learning: Unlocking Patterns for Early Intervention

Jessica Henley, Ogone Zoe Montshiwa, Joana Haushiko, Tshwetso Logic Tau

Botswana Accountancy College - Business Intelligence and Data Analytics

May 13, 2025

Declaration

We, the undersigned—Jessica Henley, Ogone Zoe Montshiwa, Tshwetso Logic Tau, and Joana Haushiko—hereby declare that this research report is our own original work. All sources of information used have been properly cited and acknowledged.

Acknowledgements

We would like to thank our supervisors, peers, and family for their support and guidance throughout this project.

Abstract

Schizophrenia is still a common and incapacitating mental illness that dramatically reduces lives around the world. Because schizophrenia often has a delayed diagnosis and extremely debilitating symptoms, there has been little development in early identification. Recent advances in data collection and artificial intelligence present an opportunity to uncover patterns that traditional approaches may overlook. This project explores machine learning predictive models to enhance early intervention by showing how proactive detection techniques and data science methodologies might enhance mental health outcomes and supplement conventional treatment procedures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of the Study

Schizophrenia is a severe, long-lasting mental illness that alters a person's thoughts, feelings, and behaviour. Because of its early onset, long-term impairment, and social burden, it has a substantial influence on public health while being relatively uncommon. According to estimates from the World Health Organization (WHO), schizophrenia affects over 24 million individuals globally, but because of its complicated and variable appearance, diagnosis and management frequently happen later.

By examining a dataset obtained from Kaggle, this study seeks to investigate how machine learning may be used to predict schizophrenia and support early intervention. The research uses the CRISP-DM approach to provide organized and ethical guidance for data mining activities.

1.2 Problem Statement

Schizophrenia's early symptoms are vague and sometimes disregarded, making early identification challenging. Usually, the diagnosis is made after psychosis has started, which reduces the efficacy of therapy and deteriorates outcomes. Traditional approaches lack the necessary accuracy for early detection. This research aims to deploy and assess ma-

chine learning algorithms to predict schizophrenia risk based on patient data, enabling timely, focused interventions.

1.3 Aim and Objectives

Aim: Develop a machine learning model to predict people with a risk of being diagnosed with schizophrenia based on demographic, socioeconomic, and clinical factors.

Objectives:

- Apply data pre-processing by cleaning and managing any inconsistencies found.
- Analyze key risk factors associated with schizophrenia through exploratory data analysis.
- Evaluate the distribution of schizophrenia diagnosis and its relation to specific features.
- Develop and train a model that will aid in identifying individuals at high risk of developing the syndrome.
- Assess the model's performance with appropriate metrics.

1.4 Significance of the Study

The research highlights how predictive analytics might assist mental health treatment, which advances the nexus between psychiatry and data science. Particularly in environments with limited resources, the study seeks to offer a replicable foundation that data scientists and mental health practitioners may build upon. It also encourages ethical considerations when dealing with sensitive health data and improves scholarly discussion of interpretable machine learning in healthcare.

1.5 Structure of the Report

The structure of the report is as follows:

- **Chapter 1: Introduction** – Provides a background, rationale, aims & objectives, and study structure.
- **Chapter 2: Literature Review** – Brief discussion of schizophrenia and examines existing work on schizophrenia risk prediction and CRISP-DM.
- **Chapter 3: Research Methodology** – Details the research design, data sources, and analytical approaches.
- **Chapter 4: Exploratory Data Analysis** – Presents data-driven visual insights and statistical summaries of the dataset.
- **Chapter 5: Results and Analysis** – Presents the experimental findings from predictive modeling, including model performance and validation.
- **Chapter 6: Discussion** – Interprets the results, examines theoretical and practical implications, and discusses limitations.
- **Chapter 7: Conclusion and Recommendations** – Summarizes conclusions, acknowledges study limitations, and offers recommendations and future research directions.

Chapter 2

Literature Review

2.1 Introduction

This chapter establishes a scholarly framework for the study of schizophrenia prediction and intervention. A literature review systematically examines previous studies and theoretical models relevant to using machine learning for early schizophrenia detection. In this instance, it contextualizes the use of data mining frameworks like CRISP-DM and the development of predictive models. The review highlights key concepts such as risk prediction, symptom categorization, and algorithmic transparency.

By synthesizing best practices and lessons from earlier research, this chapter strengthens the study's methodological choices and situates the current work within the broader academic context of data-driven mental health care.

2.2 Understanding Schizophrenia: A Global Health Concern

Approximately 1% of people worldwide suffer from schizophrenia (Stilo and Murray, 2010). According to (Lewis and Lieberman, 2000), schizophrenia is a widely misdiagnosed and stigmatised mental disorder that manifests as "abnormal mental functions and

disturbed behaviour."

Under the guise of different names, schizophrenia has been tormenting mankind since the beginning of their first existence. Only in the late 18th century and early 19th century was a correlation made of symptoms from seemingly separate illnesses coined together to belong to just one, schizophrenia. Further studies and research into the sickness to date has enhanced diagnosis, treatment and quality of life but has not produced a cure, with only a small proportion of patients falling into remission.

2.2.1 A Short History of Schizophrenia

Throughout history, schizophrenia has been identified in a variety of ways. As early as 2000 BC, ancient Egyptian, Hindu, and Chinese texts mention symptoms that are currently linked to the disease. Exorcisms and lobotomies were among the treatments that resulted from early interpretations that suggested supernatural possession or divine punishment.

Hippocrates was one among the first to question these theories, claiming that the illness was caused by brain imbalances. Emil Kraepelin officially recognised it as dementia praecox in the 19th century. The term'schizophrenia' was later created by Eugen Bleuler in 1911, who distinguished between positive and negative symptoms and emphasised the separation of cognition and emotion.

The advent of antipsychotic drugs in the 1950s marked a tremendous advancement in medical knowledge and care. Although there is currently no known cure for schizophrenia, it is controlled with a mix of medication, counselling, and psychological support.

(The entirety of this information was learned from (Chr Kyziridis, no date))

It is important to look back and understand where we came from, what worked and what didn't to better understand and build on how to best move forward.

2.2.2 Stages of Schizophrenia

According to (McGorry et al., 2008), schizophrenia develops in four stages. Firstly, risk followed by prodrome then psychosis and finally chronic disability.

Risk is the possibility that a person will be diagnosed with schizophrenia depending on environmental and preset factors.

1. **Perinatal (between pregnancy and one year after delivery) and prenatal (during pregnancy) Conditions**

   Particularly, the findings showing a marginally elevated incidence of psychosis in children who experienced prenatal damage, maternal malnourishment, second-trimester infections, traumatic insults, and stress during infancy are covered by (Insel, 2010) and (Lewis and Lieberman, 2000).

   This is further supported by (Kahn et al., 2015), also pointing out that among patients with schizophrenia, those born in the spring and late winter were somewhat over-represented. This is most likely because of respiratory infections in the mother or wintertime malnutrition (such as vitamin D deficiency).

   All three papers however do establish that no definitive grounds back these claims, and they are but mild effects.

2. **Paternal Age**

   (Malaspina et al., no date) talks about the possible consequences of mutations linked to paternal age. It was discovered that males who fathered children later in life were more likely to have a child with schizophrenia. It was unclear, nevertheless, if this was because of biological or psychological factors.

3. **Sex**

   According to (McGrath et al., 2008), there are almost 1.4 times as many male patients as female patients. Compared to women, men are often more likely to get the syndrome. Additionally, they exhibit symptoms earlier than women, usually in

their early 20s as opposed to women's late 20s.

4. **Genetic Factors**

   People are more at risk if they have family members or relatives who have had a prior diagnosis. People with one schizophrenic parent had an approximate 10% chance of getting the illness, whereas those with two schizophrenic parents had an astounding 40% chance, according to a study by (Kringlen, 2000).

5. **Drug Abuse**

   According to (Hamilton, 2017) and (Vaucher et al., 2018), regular cannabis usage in the early adolescent years was linked to greater positive symptoms like hallucinations. This does not imply that all marijuana users will develop schizophrenia; rather, it suggests that if a person has a certain gene and smokes, the two may combine to cause the illness. In 2014, McLoughlin et al. conducted a review which revealed that 40% of patients with schizophrenia use cannabis.

6. **Urban Environment**

   In order to determine where the majority of schizophrenia patients came from, (Kahn et al., 2015) examined a number of studies carried out in both rural and urban locations in an effort to find a common theme that might be considered a risk factor. Large cities mostly produced the most patients, according to the findings of all these investigations.

7. **Migration**

   Moving to a new country was found to play "a role for psychosocial adversity in the aetiology of schizophrenia" by (Cantor-Graae and Selten, 2005), who analysed several studies and synthesised the data gathered on migration.

8. **Social Hardships**

   Adversities in childhood and adulthood that have been identified as trigger events that contributed to the disease's emergence are covered in (Stilo and Murray, 2010).

These included parental separation or death, bullying, and sexual, physical, or mental abuse. Additionally, research indicated that a relapse of the illness would be preceded by severely stressful situations.

These are the most common and fairly definitive risk factors discussed and agreed upon by many studies and articles. There is a high chance that there could be more unidentified risk factors that contribute to developing schizophrenia. These risks are important to know as they directly relate to building the model.

The next stage is prodrome. The prodrome phase is "identified based on changes in thoughts (for example, bizarre ideas falling short of psychotic ideation), social isolation and impaired functioning (for example, reduced school performance)"(Insel, 2010).

Psychosis manifests after the prodrome phase. This is when the disease becomes the most apparent with full-blown delusions and hallucinations. Majority of times, it is at this point, when things have escalated to an extreme, that the disease is then identified.

Chronic disability is when a person's ability to work independently in their day-to-day life is impaired. While symptoms can be managed, normal activities healthy individuals find easy to do (e.g. self-care routines, maintaining healthy relationships, etc.) can become extremely challenging for schizophrenic patients due to cognitive impairments and symptoms like hallucinations.

These are the 4 stages of schizophrenia. So far, no cures for the illness have been discovered, therefore people who have been diagnosed with schizophrenia, unfortunately have to live with it for the rest of their lives.

What this project is trying to achieve is to predict individuals who carry the risk of developing schizophrenia, in order to bring about early interventions which could potentially aid them to better manage their symptoms.

Data-driven mental health research is gaining popularity due to the need for innovative solutions. Machine learning has exciting opportunities to support clinical practice by detecting risk patterns and forecasting disease onset, thanks to large datasets and advances

in computational techniques.

2.3 Framework Establishment: CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a well-known six-phase framework for organizing data-driven projects (Shearer, 2000). Its phases are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Chapman et al., 2000). CRISP-DM is industry-neutral and provides a structured, iterative process that aligns technical work with business and domain objectives. This makes it particularly suitable for healthcare applications, where ethical and domain-specific considerations must be integrated at every step. By following CRISP-DM, this study ensures that each stage of analysis—from feature selection to model validation—adheres to best practices in data mining.

2.4 Existing Literature on Machine Learning in Schizophrenia Research

Machine learning (ML), which offers potential paths for predictive modeling and pattern identification, has become a game-changing tool in the early diagnosis and treatment of schizophrenia.

According to a study by (Vellido et al., 2019), a variety of algorithms that have been widely used in psychiatric contexts for classification tasks include support vector machines (SVM), decision trees, and random forests. Logistical regression, support vector machines, random forests, and ensemble methods have shown potential in predicting schizophrenia diagnoses using clinical and non-clinical data (Orrù et al., 2020). According to (Mørch et al., 2018), these models have proven to be very accurate, especially in research using behavioural and organized neuroimaging datasets. In another instance, Koutsouleris et al. (2018) employed ML models to predict patients at clinical high risk and achieved good predictive performance.

However, in clinical settings where decision-making need understandable insights, its implementation frequently prioritizes predictive performance above interpretability and transparency, two crucial drawbacks.

The requirement for interpretable models is becoming more widely recognized in emerging literature, especially in practical healthcare research. Because logistic regression is transparent and simple to use, recent research has investigated its potential for predicting symptom intensity, adherence behaviour, and treatment results (Sharma et al., 2021). Although logistic regression is not as complex and powerful as ensemble or deep learning models, it provides a clear view of the impact of individual variables (Shafiei et al., 2021), which is crucial when results are meant for policymakers and doctors. Even in large datasets, complex models like SVMs and XGBoost showed no major advantage over logistic regression with an average AUC score of about 0.70 (Bracher-Smith et al., 2022). To further prove it's good performance, in a remission prediction study, logistic regression performed well on real clinical data and was externally validated in Canada, showing practical reliability (Shafiei et al., 2021). Logistic regression has been validated in real-world settings.

Even though initial research was highly reliant on neuroimaging or cognitive testing, there has been a move towards utilizing electronic health records (EHRs) and demographic information, expanding the scope of ML applications. Such models facilitate quicker, earlier, and more convenient pathways to diagnosis, particularly in low-resource environments.

Few studies specifically include the CRISP-DM paradigm into machine learning research pertaining to schizophrenia; instead, ad hoc processes are frequently employed. As a result of this lack of methodological consistency, data preparation, model evaluation, and interpretability procedures have become inconsistent. Using structural neuroimaging data, (Martínez-Murcia et al., 2020) used deep learning to classify schizophrenia; nevertheless, their "black-box" model's opacity and scant attention to deployment feasibility drew criticism. Strong and clinically useful results are typically obtained from

studies that follow more closely to CRISP-DM principles, such as incorporating domain knowledge during business understanding and guaranteeing iterative review.

In this regard, the current study prioritizes methodological rigor and interpretability within the CRISP-DM framework rather than aiming for pure predictive performance alone to test its methodology against previous literature. This intentional alignment guarantees that every stage of the study, from feature selection to deployment, is methodically based on data science best practices and clinical relevance. This alignment is supported by the decision to concentrate on interpretable techniques, such as **logistic regression**, which improve transparency and make it easier for data scientists and mental health specialists to communicate across disciplines.

2.5 Summary

It is clear from combining the strands of research discussed that although many advanced machine learning models have high predictive accuracy, problems with non-transparency and complexity usually prevent them from being used in clinical settings. In contrast, the present study purposefully uses logistic regression within a CRISP-DM framework in order to prioritise interpretability. This method not only guarantees that all phases—from data preparation to model deployment—are grounded in methodological best practices, but it also makes it easier for people from different disciplines to communicate openly. To ensure that predictive models for schizophrenia not only predict psychosis early but are also sufficiently clear for routine clinical usage. Future research should continue examining the balance between model complexity and interpretability.

Chapter 3

Research Methodology

This chapter describes the methodological process undertaken to develop a predictive model to predict schizophrenia risk.

3.1 CRISP-DM

The study utilizes the Cross-Industry Standard Process for Data Mining (CRISP-DM), a widely adopted data science framework that facilitates systematic and reproducible analysis. This structure provided a clear path for the project so that each phase was based on best practice and ethical standards, particularly important in dealing with sensitive health data.

1. Business Understanding

The investigation began with business understanding, in which the overarching goal was determined: to produce an understandable, unbiased, and accurate predictive model that is capable of identifying persons at risk for schizophrenia based on demographic, clinical, and psychosocial factors. The goal was to support early diagnosis and inform treatment choices, ultimately contributing to better mental health outcomes. This phase determined the scientific and practical value of the project, ensuring alignment with healthcare demand and the ethical requirement for expanded diagnostic access and equity.

2. Data Understanding

Understanding data followed, in which reviewing the composition and characteristics of the dataset that was downloaded from Kaggle was performed. The information included 10,000 anonymized patient records with 20 features like gender, age, mental illness family history, compliance with medication, education level, stress levels, and Global Assessment of Functioning (GAF) scores. Initial inspection showed that the data was primarily clean with few missing values, no outliers and class imbalance. The original dataset was in Turkish and was given in an unstructured format. This required translation and contextual interpretation to make the data meaningful and analytically useful. Lack of contextual metadata and full transparency about how data were collected are shortcomings of secondary sources, but these were anticipated through careful scrutiny and standardization.

3. Data Preparation

In the data preparation phase, the dataset was cleaned and transformed with a combination of Excel and Python. Translating was done to convert Turkish entries into English, after which the dataset was rearranged in a tabular format. Categorical binary fields such as gender and diagnosis status were converted into numerical equivalents (0 and 1) for simplicity in machine learning modeling. Missing values were filled out where needed. Categorical variables were further standardized to create uniformity among entries. Feature selection was guided by domain knowledge and insights from exploratory data analysis (EDA) to retain the most significant predictors of schizophrenia for modelling. This rigorous preparation enhanced data quality and ensured robustness of subsequent predictive operations. Relevant features were selected based on domain relevance and EDA findings.

4. Modelling

Modelling involved surveying several machine learning models, including logistic regression and random forest. Model selection was based on a balance between accuracy, interpretability, and clinical relevance (logistic regression was deemed fit and most relevant). During the training, the model was fine-tuned and cross-validated to prevent overfitting

and allow a certain level of generalizability to new patient data.

5. Evaluation

In the evaluation stage, the model was tested based on a variety of metrics, including precision, accuracy, recall, and F1-score. Special care was taken in the medical context while reducing false negatives because failing to recognize the risk of schizophrenia can have serious consequences.

6. Deployment

An interactive dashboard was created to visualize findings, facilitate clinical decision-making, and communicate results to stakeholders. The dashboard allows dynamic comparison of variables influencing schizophrenia risk and treatment outcomes. The dashboard enables stakeholders such as clinicians, researchers, and policymakers to see key insights and interact with the information via variables. A well-trained and well-performing model was created that could predict individuals with high risk of schizophrenia with high accuracy. It identifies these individuals based on select features. This allows for medical professionals to use it as an aid in diagnosis or prevention, and to see how each factor directly impacts risk levels.

3.2 Source of Data

The dataset used in this study came from Kaggle, a reliable online resource for structured data sharing. 10,000 anonymised patient data points with twenty characteristics, including age, gender, family history, stress levels, and medication adherence, are included.

Advantages of Using Kaggle as a Secondary Data Source

- *Open and Free Access*: Kaggle provides open and free access to a wide range of datasets in fields such as healthcare, finance, education, and social sciences
- *Large Volume and Variety of Data*: Kaggle datasets usually have big sample sizes (e.g., the 10,000 anonymised patient records in your project), which allows for

substantial statistical analysis and perhaps stable model performance.

- *Ethical Accessibility and Anonymization*: The majority of datasets on Kaggle are anonymized and ethically sourced, which is favourable to research ethics—especially in sensitive fields like healthcare.

Disadvantages of Using Kaggle as a Secondary Data Source

- *Limited Control Over Data Collection*: Researchers lack control over data collection of how variables were or were not measured, or under what conditions data were recorded—decreasing flexibility and depth of analysis.
- *Limited Contextual Information*: While most datasets are organized, they may lack background context or critical qualitative information. This may limit interpretation or lead analysis astray when underlying assumptions are not known.
- *Duplication and Redundancy*: Some Kaggle datasets are duplicate copies or reuploads of existing public datasets, and this can lead to redundancy if not efficiently verified.

3.3 Data Analysis and Evaluation Overview

We began by cleaning the dataset in Excel, imputing missing values and scaling categorical fields like diagnosis and gender. Exploratory Data Analysis (EDA) was conducted in Python using Pandas, Matplotlib, and Seaborn to identify patterns in primary variables like schizophrenia diagnosis, gender, area they reside in, family history, medication adherence, GAF scores, and education.

Visualizations like bar plots and distribution charts indicated class imbalances, symptom trends, and adherence effects. An interactive Excel dashboard was developed to enable dynamic comparisons across variables, which aided interpretation and informed predictive model construction.

3.3.1 Evaluation of Data Quality

The schizophrenia dataset used for this project was generally well-prepared and relatively clean. No duplicate entries, and consistent data types. However, one key issue identified during data quality evaluation was class imbalance, where the number of diagnosed schizophrenia cases was significantly lower than non-diagnosed cases. This imbalance posed a challenge for predictive modelling, as machine learning algorithms tend to be biased toward the majority class, leading to poor detection of minority (positive). Additionally, the dataset was originally in Turkish, requiring translation into English to ensure consistency and interpretability. Categorical binary fields were also converted into numerical values (0s and 1s).

3.4 Ethical Considerations

This study employed anonymized secondary data collected from Kaggle such that there is no personally identifiable information utilized. Ethical approval was not required as the dataset is de-identified and open to the public. However, adequate data handling procedures were put in place for confidentiality and privacy controls.

All the results were interpreted responsibly to avoid stigmatizing patients with schizophrenia.

In summary, this methodology employed the CRISP-DM approach to guide the development of a predictive model for schizophrenia risk. Each phase—from data gathering and preprocessing to modeling and deployment—was conducted with a high level of data quality, clinical usefulness, and ethical responsibility. This rigorous process ensures the final model is not only scientifically valid but deployable and consistent in real-world mental health settings.

Chapter 4

Exploratory Data Analysis

4.1 Overview

The exploratory data analysis (EDA) aims to uncover patterns, trends, and correlations within the schizophrenia dataset. Before model training, EDA improves understanding of feature distributions, data quality, and interactions. This phase provides insights on outliers, missing values, and multivariate relationships pertinent to schizophrenia diagnosis and treatment adherence.

4.2 Distribution of Schizophrenia Diagnosis

Figure 4.1 shows the distribution of schizophrenia diagnoses in the dataset. Of 10,000 people, 2,887 (28.87%) had a schizophrenia diagnosis, whereas 7,113 (71.13%) did not. This class imbalance is consistent with epidemiological expectations and highlights the need for careful handling (e.g., class weighting or resampling) to avoid biasing models toward the majority class.

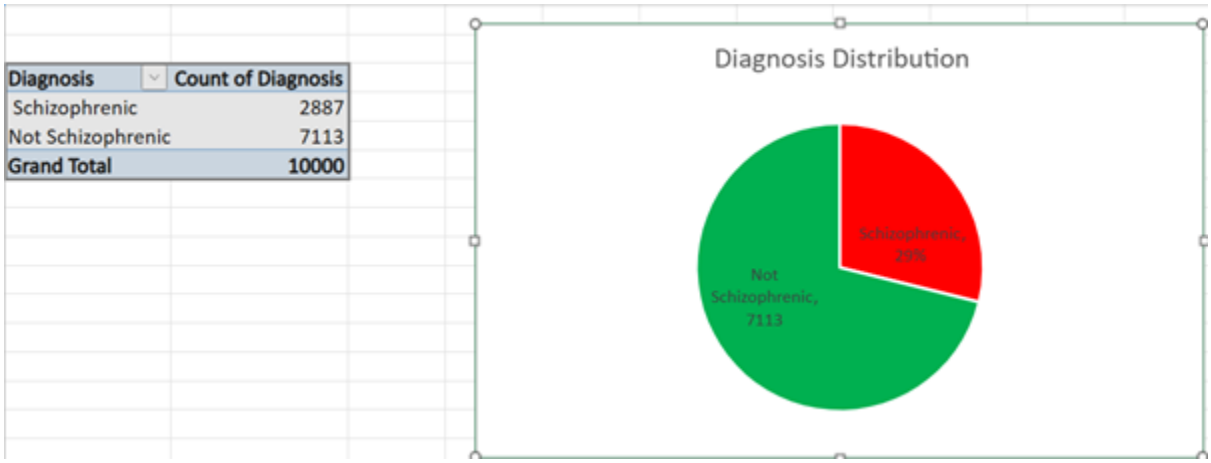

Figure 4.1: Distribution of schizophrenia diagnoses in the dataset.

4.3 Gender and Schizophrenia Diagnosis

Gender can influence schizophrenia presentation. In the dataset, males and females are equally represented. Analysis shows a slight gender difference: men have a slightly higher diagnosis rate than women, while women make up a larger share of non-cases. This aligns with research indicating earlier onset and greater severity in males. Figure 4.2 illustrates the gender-based distribution of schizophrenia diagnosis.

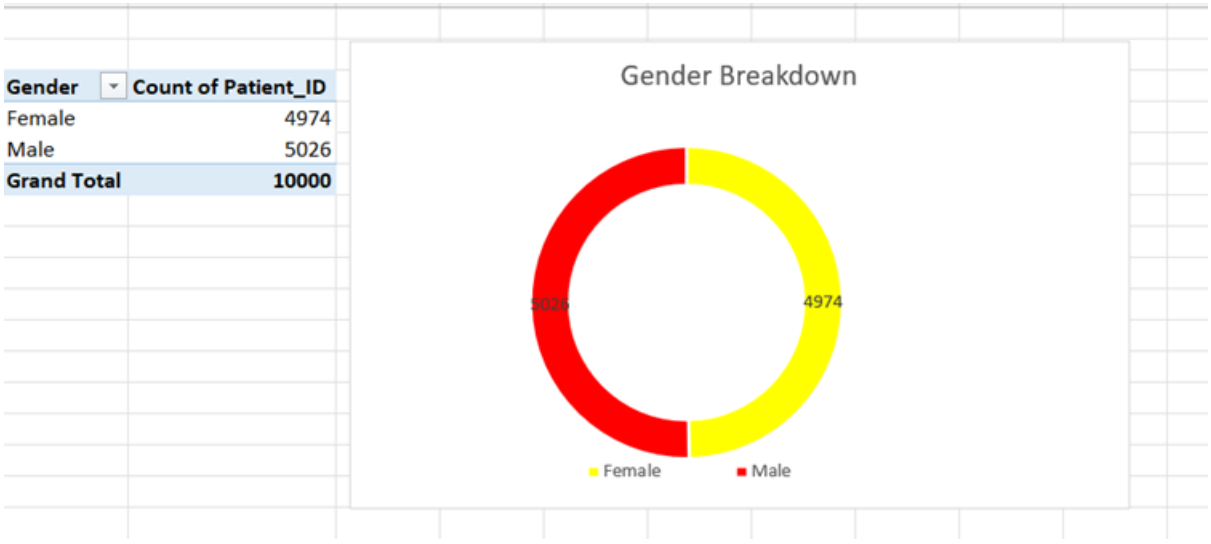

Figure 4.2: Gender-based distribution of schizophrenia diagnosis.

4.4 Living Area and Schizophrenia Diagnosis

Urban versus rural residence may affect schizophrenia prevalence. Figure 4.3 compares diagnoses by living area. In urban regions, 1,499 out of 5,006 people (29.93%) were diagnosed, whereas in rural areas, 1,388 out of 4,994 (27.80%) were diagnosed. This confirms a slight urban bias.

Count of Patient_ID	Column Labels		
Living Area	Schizophrenic	Not Schizophrenic	Grand Total
Rural	1388	3606	4994
Urban	1499	3507	5006
Grand Total	**2887**	**7113**	**10000**

Figure 4.3: Comparison of schizophrenia diagnosis by living area.

4.5 Family History and Schizophrenia Diagnosis

Family history is a strong risk factor. In the dataset, 1,748 of 3,196 individuals with a family history (54.7%) were diagnosed, compared to 1,139 of 6,804 without history (16.7%). Figure 4.4 shows the diagnostic rates with/without family history, indicating a substantially higher risk for those with genetic predisposition.

Count of Patient_ID	Column Labels		
Family History	Schizophrenic	Not Schizophrenic	Grand Total
No	1139	5665	6804
yes(Generic predisposition)	1748	1448	3196
Grand Total	**2887**	**7113**	**10000**

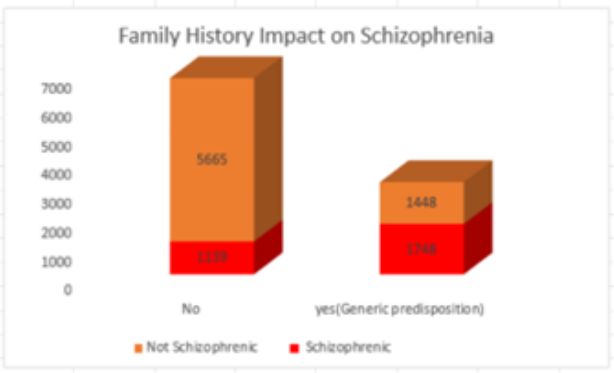

Figure 4.4: Impact of family history on schizophrenia diagnosis.

4.6 Medication Adherence and Suicide Attempts

Medication adherence is critical. Figure 4.5 shows the proportion of patients who attempted suicide by adherence level. Suicide attempts rise sharply as adherence declines, illustrating the protective effect of consistent treatment.

Count of Patient_ID	Suicide Attempt		
Medication Adherence	No	Yes	Grand Total
Good	4001	154	4155
Moderate	2796	269	3065
Poor	2319	461	2780
Grand Total	**9116**	**884**	**10000**

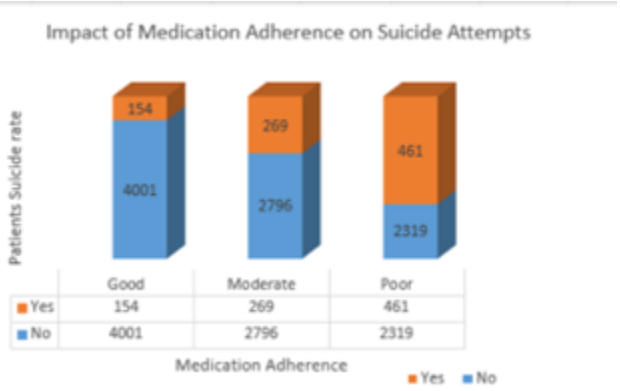

Figure 4.5: Suicide attempts by medication adherence level.

Key breakdown:

- Good adherence (4,155 patients): 154 (3.7%) attempted suicide.
- Moderate adherence (3,065 patients): 269 (8.8%) attempted.
- Poor adherence (2,780 patients): 461 (16.6%) attempted.

This trend underscores that lower adherence is associated with higher suicide risk.

4.7 Medication Adherence and Positive Symptom Severity

Figure 4.6 shows average positive symptom scores by adherence level. Lower adherence is linked to higher symptom severity, aligning with clinical knowledge that inconsistent medication use exacerbates positive symptoms.

Medication Adherence	Average of Positive symptom score
Good	31.23995187
Moderate	39.1497553
Poor	50.50143885
Grand Total	39.019

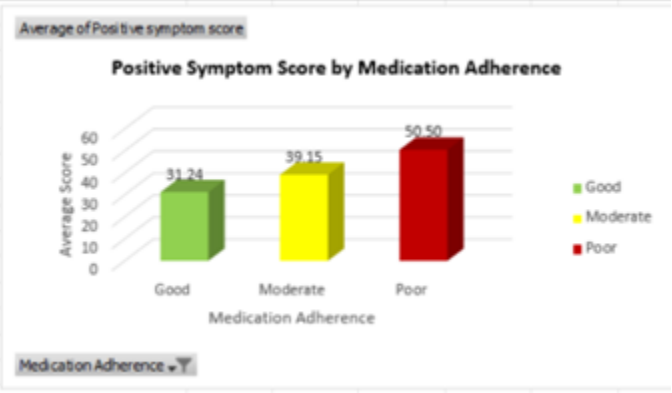

Figure 4.6: Mean positive symptom score by medication adherence level.

4.8 Average GAF Score by Disease Duration

Figure 4.7 displays average Global Assessment of Functioning (GAF) scores by disease duration.

Row Labels	Schizophrenic	Not Schizophrenic	Grand Total
0		79.98495712	79.98495712
1	35.14473684		35.14473684
2	35.12643678		35.12643678
3	39.25925926		39.25925926
4	34.06756757		34.06756757
5	36.4057971		36.4057971
6	35.82857143		35.82857143
7	32.86764706		32.86764706
8	35.96363636		35.96363636
9	34.57352941		34.57352941
10	37.94117647		37.94117647
11	34.94736842		34.94736842
12	34.25		34.25
13	34.05063291		34.05063291
14	32.3902439		32.3902439
15	35.08450704		35.08450704
16	37.59722222		37.59722222
17	34.97435897		34.97435897
18	35.13333333		35.13333333
19	34.8045977		34.8045977
20	32.82432432		32.82432432
21	32.44		32.44
22	33.88732394		33.88732394

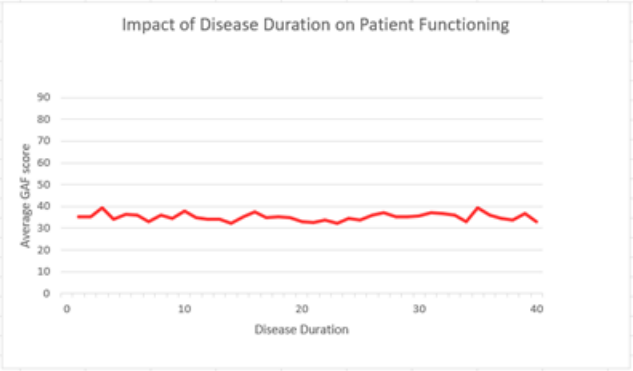

Figure 4.7: Average GAF score by medication adherence level.

Key Findings:

- Schizophrenic patients' GAF scores remain relatively flat, averaging around 35, regardless of how long they've had the illness.
- Unlike in cases of good medication adherence where GAF can climb to 74, here we see no sign of recovery or improvement with time.

- Even patients with 30+ years of disease history don't cross into higher functioning scores, showing prolonged stagnation.

This pattern suggests either chronic non-adherence, treatment resistance, or both.

4.9 Educational Attainment Among Schizophrenia Patients

Educational level can reflect socioeconomic and cognitive factors. Figure 4.9 shows the distribution of educational attainment. The categories are fairly balanced, with post-graduate education being the largest group (21.43%). This suggests schizophrenia affects people across all educational levels.

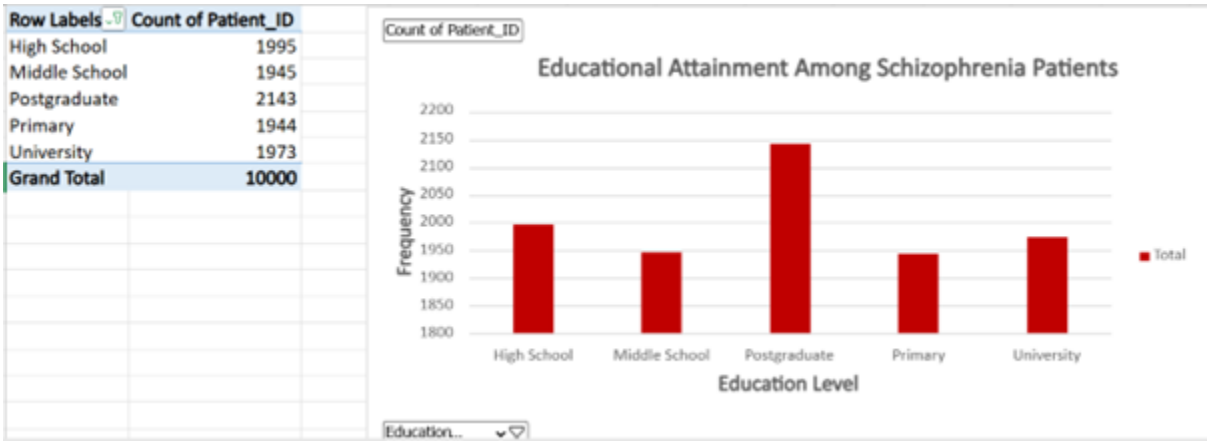

Figure 4.8: Distribution of educational attainment among schizophrenia patients.

Key Findings:

- Postgraduate level is the largest group among schizophrenia patients, with 599 cases (out of a total 2,143 patients, or 21.43%).
- High School follows closely at 568 cases (1,995 total; 19.95%).
- University educated patients number 598 (1,973 total; 19.73%).
- Middle School comes next with 562 cases (1,945 total; 19.45%).

- Primary School has 560 cases (1,944 total; 19.44%).

4.10 Dashboard

After editing and changing of the colour schemes, a more neater and visually appealing dashboard was created using the same information on the graphs:

Figure 4.9: Complete View of Dashboard

Chapter 5

Results and Analysis

5.1 Introduction

This chapter provides an overview of the experimental results from the investigation into using an interpretable logistic regression model for early schizophrenia risk prediction. The study is based on a thorough five-fold cross-validation process followed by a final assessment on a test set that has been held out. The chapter outlines the improvements made from early iterations that produced unrealistically perfect scores, explains significant changes in data preprocessing and model structure, and includes numerical summaries and a loss curve visualisation to support our findings to address the research questions and objectives.

5.2 Analysis and Presentation of Data

5.2.1 Enhancements and Model Building

The model generated nearly faultless performance indicators at the beginning of the investigation. Data leakage was the main cause of these deceptively excellent ratings, according to analysis; numerical values for categorical categories (such "Family History") were mistakenly handled as continuous variables.

In order to fix this, continuous and categorical features were separated in the preprocessing pipeline and one-hot encoding was used to maintain the discrete nature of the latter. This stage made sure that every input characteristic would clearly contribute to the final prediction and prevented the model from catching false patterns. As a result, subsequent model iterations showed more robust and somewhat more realistic results throughout the five-fold cross-validation procedure.

Using the **Keras Functional API**, a more transparent logistic regression model that takes individual inputs for each feature was created. The probability is output by feeding each feature via its input layer, concatenating it afterwards, and then passing it to a dense layer with a sigmoid activation. The RMSprop optimizer is used to construct the model, and measures like accuracy, precision, recall, and AUC are used to assess it. Key code components include

- **Input Creation:**

```
model_inputs = [Input(name=name, shape=(1,)) for name in input_feature_names]
```

- **Feature Concatenation:**

```
concatenated_inputs = layers.Concatenate()(model_inputs)
```

- **Output Definition and Compilation:**

```
output = layers.Dense(units=1, activation="sigmoid")(concatenated_inputs)
model = Model(inputs=model_inputs, outputs=output)

model.compile(
    optimizer=optimizers.RMSprop(learning_rate),
    loss=losses.BinaryCrossentropy(),
    metrics=[
        metrics.BinaryAccuracy(name='accuracy', threshold=classification_threshold),
        metrics.Precision(name='precision', thresholds=classification_threshold),
        metrics.Recall(name='recall', thresholds=classification_threshold),
        metrics.AUC(num_thresholds=100, name='auc')
    ]
)
```

With this modular approach, every input characteristic is guaranteed to retain its integrity and visibly contribute to the final prediction. In this project, NumPy was used to create a logistic regression model from scratch rather than depending on pre-built logistic regression implementations from libraries like scikit-learn. This method provides total control over each algorithmic step, which made it simpler to comprehend and adapt the model to the study's unique requirements for early schizophrenia risk prediction.

5.2.2 Five-fold Cross Validation and Test Set Evaluation

To more accurately assess the logistic regression model's performance, 5-fold cross-validation was used. By dividing the dataset into five equal parts, this technique aids in lowering bias and variance in the model evaluation. The model is trained on four parts and tested on the remaining one, and this process is repeated five times—each fold serving as the test set once. A more precise assessment of the model's generalisation performance, particularly when dealing with sparse data, can be obtained by averaging these outcomes.

Table 5.1 shows cross-validation metrics (loss, accuracy, precision, recall, AUC) for each fold.

Fold	Loss	Accuracy	Precision	Recall	AUC
1	0.07370	0.9900	0.9711	0.9958	0.9998
2	0.09777	0.9762	0.9421	0.9785	0.9989
3	0.08879	0.9856	0.9670	0.9821	0.9993
4	0.13331	0.9781	0.9494	0.9761	0.9987
5	0.07961	0.9869	0.9662	0.9892	0.9996
Average	0.09464	0.9834	0.9592	0.9844	0.9992

Table 5.1: 5-Fold Cross-Validation Metrics for the final model.

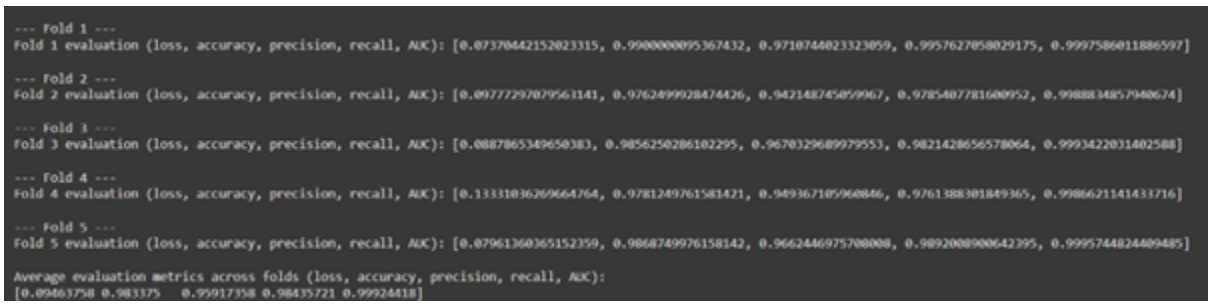

Figure 5.1: Code Result for 5-Fold Cross-Validation

The model achieved an average accuracy of 98.34%, precision of 95% and AUC of 0.9992.

For final evaluation, a new model was trained on the combined training data and tested on 1,000 holdout samples.

Performance was:

- Loss = 0.05103
- Accuracy = 0.9870
- Precision = 0.9630
- Recall = 0.9931
- AUC = 0.99978

The confusion matrix (Table 5.2) shows very few errors.

	Predicted Neg.	**Predicted Pos.**
Actual Neg.	701	11
Actual Pos.	2	286

Table 5.2: Confusion matrix for test set predictions.

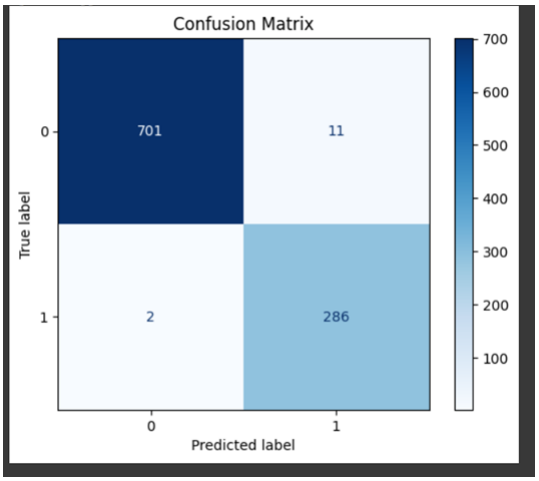

Figure 5.2: Confusion Matrix

5.2.3 Loss Curve Insights

The loss curve is a crucial sign of training stability. Plotting the training loss against epochs, the loss curve exhibits a gradual decrease until leveling out. Without displaying unpredictable behaviour or indications of overfitting during training, this smooth convergence verifies that the model is learning as intended.

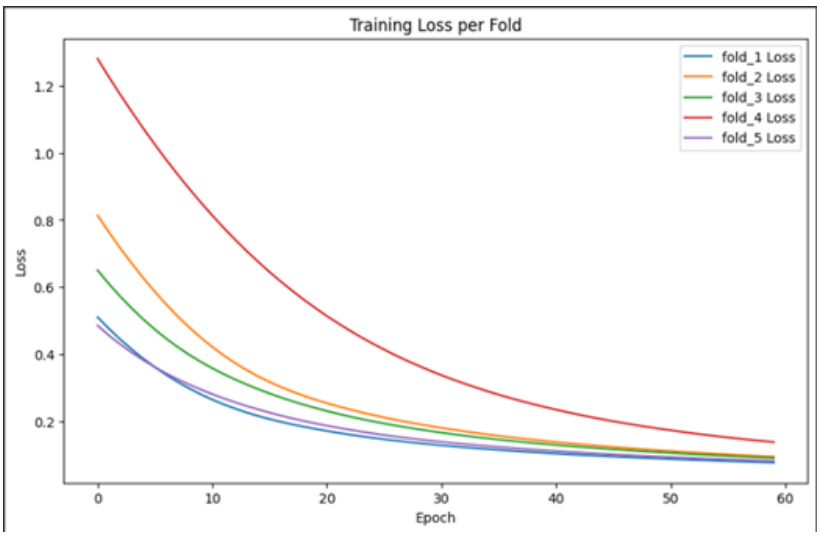

Figure 5.3: Loss Curve

5.3 Conclusions

The results confirm that the updated model, which is improved by appropriate preprocessing (especially for numerical categorical variables) and a more careful architectural design, consistently performs well in terms of prediction. The logistic regression model demonstrates both reliability and generalizability for early schizophrenia risk prediction, with a robust test set evaluation and an average cross-validation accuracy of 98.34%.

Chapter 6

Discussion

6.1 Synopsis of the Study Results

The goal of the study was to use clinical and demographic data to create an interpretable model that would predict the risk of schizophrenia. These results, which came from a final test set evaluation and five-fold cross-validation, show that the model performs remarkably well, with a precision score of over 96% and an AUC of about 1.0. The findings support the idea that an interpretable logistic regression model can be a useful tool for early intervention in schizophrenia if it is appropriately pre-processed and thoroughly cross-validated.

6.2 Theory and Practice Implications

Key implications:

- **Theoretical:** Demonstrates that careful feature encoding and preprocessing allow simple models to achieve high accuracy in clinical prediction, contributing to the literature on explainable models in psychiatry.
- **Practical:** Supports the use of such interpretable models in clinical decision support, as they maintain transparency and high performance, potentially improving

early detection of schizophrenia in practice.

6.3 Unexpected Outcomes and Limitations

The model generated nearly flawless results in the early tests, but they were ultimately found to be misleading. The main cause of these outcomes was data leaking. This was the consequence of an original simple preparation pipeline that normalized all numerical columns, including those representing category data.

By separating continuous and categorical variables and using a single hot encoding for the latter, these problems were resolved, significantly enhancing the generalizability of the model and yielding more accurate performance measurements.

Another potential reason could be that ten thousand anonymised patient records with minimal missing values, no outliers, and a controllable class imbalance were included in the Kaggle dataset. These circumstances most likely enhanced learning effectiveness and model stability.

Because the dataset came from a single source, it lacked demographic diversity. It's possible that this homogeneity produced unduly positive findings that might not apply to more diverse groups. Furthermore, no external validation—which is necessary to verify robustness—was carried out.

Even though the final model performs exceptionally well, care should be taken to account for any biases in the dataset and the necessity of external validation to support these results.

Chapter 7

Conclusions and Recommendations

7.1 Conclusions

This study shows that schizophrenia risk could potentially be accurately and reliably predicted by a well-built logistic regression model. After extensive preprocessing and five-fold cross validation, the final model produced nearly flawless AUC values and a precision of 96% on the test set. These findings support the viability of early intervention systems built on interpretable analytics and the need of appropriate data management in preventing inaccurate performance predictions.

Compared to logistic regression models from other studies like (Bracher-Smith et al., 2022) with an AUC score of about 0.70, this project's model outperformed theirs, this could be due to varying feature variables and datasets. The data that was used in those model trainings may not have been as clean and uniform as the dataset used in this study, it may have been more representative of real-world data.

All in all, this study was able to achieve its overall aim and objectives, despite hitting a few roadblocks.

7.2 Study Limitations

Limitations include:

- **Dataset Limitations:** There is little demographic variation in the dataset utilised in this study because it was gathered from a single source. The generalisability of the findings may be impacted by this lack of heterogeneity. For instance, important variables that are impacted by socioeconomic, cultural, or regional factors may not be adequately represented. Therefore, it's possible that the claimed very high performance measures don't apply to populations with other features. The use of more representative and varied datasets might be advantageous for future research.
- **Model Specificity:** Certain features of the dataset may be partially responsible for the model's excellent performance. The performance measures may have been unintentionally overstated due to biases or special characteristics in the data. Therefore, to verify the model's robustness and make sure that the observed performance is not unduly adapted to the sample at hand, it is advised to carry out additional assessments using external data.

7.3 Recommendations

In light of the research conducted, the following in advised:

- **For Practitioners:** Consider integrating interpretable prediction models into clinical processes for the early identification of schizophrenia, while keeping in mind that thorough preprocessing is essential.
- **For Policy Makers:** Policymakers should encourage and support cooperative initiatives that make it easier to gather more varied and comprehensive information in order to increase the accuracy and relevance of predictive models in mental health. By encouraging more participation across various demographics and geographical areas, such programs can improve the generalisability of upcoming models.

- **Future Research:** To further improve transparency, future research could investigate adding more clinical characteristics, use external validation to assess the model's resilience (testing the model on independent datasets), and use Explainable AI (XAI) techniques.

7.4 Future Research Directions

Future research should focus on the following areas in order to expand on the current study and enhance early intervention techniques for schizophrenia:

External Validation: Use different datasets that span a wider range of demographics to confirm the model's performance. In order to preserve the model's high accuracy and robustness across many populations and environments, this phase is essential.

Examining Advanced Models: While maintaining interpretability, look at the application of ensemble techniques or more complex models. Although sophisticated methods might improve prediction accuracy, they should be weighed against the requirement for openness so that physicians can comprehend and have confidence in the forecasts.

Longitudinal Data Analysis: Studies that follow patients over time can be used to identify dynamic trends in risk variables and the development of symptoms. More proactive and specialised intervention techniques could be informed by longitudinal analysis, which could offer greater insights into how schizophrenia risk changes over time.

Bibliography

[1] Martínez-Murcia, F. J., Górriz, J. M., Ramírez, J., Salas-Gonzalez, D., & Segovia, F. (2020). A deep learning approach for early diagnosis of schizophrenia based on structural MRI data. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 100, 109913. https://doi.org/10.1016/j.pnpbp.2020.109913

[2] Mørch, N., Lykke, J., & Rask, M. T. (2018). Machine learning techniques in the prediction of schizophrenia—a systematic review. *International Journal of Methods in Psychiatric Research*, 27(3), e1725. https://doi.org/10.1002/mpr.1725

[3] Sharma, P., Chaudhary, H., & Bansal, D. (2021). Predictive modeling in schizophrenia using linear regression: Toward interpretable clinical tools. *Journal of Psychiatric Research*, 139, 269–278. https://doi.org/10.1016/j.jpsychires.2021.05.036

[4] Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13–22.

[5] Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2019). Making machine learning models interpretable for clinical psychiatry. *Computer Methods and Programs in Biomedicine*, 180, 105020. https://doi.org/10.1016/j.cmpb.2019.105020

[6] Koutsouleris, N., et al. (2018). Prediction models of functional outcomes for individuals at clinical high risk for psychosis: A multimodal, multicenter machine learning approach. *JAMA Psychiatry*, 75(4), 398–406.

[7] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

[8] Bracher-Smith, M. et al. (2022). "Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank," *Schizophrenia Research*, 246, pp. 156–164.

[9] Cantor-Graae, E. and Selten, J.-P. (2005). Schizophrenia and Migration: A Meta-Analysis and Review. *Am J Psychiatry*.

[10] Hamilton, I. (2017). "Cannabis, psychosis and schizophrenia: unravelling a complex interaction," *Addiction*, 112(9), pp. 1653–1657.

[11] Insel, T. R. (2010). "Rethinking schizophrenia," *Nature*, pp. 187–193.

[12] Kahn, R. S. et al. (2015). "Schizophrenia," *Nature Reviews Disease Primers*, 1.

[13] Kringlen, E. (2000). "Twin studies in schizophrenia with special emphasis on concordance figures," *Am. J. Med. Genet.*, 97(1), pp. 4–11.

[14] Lewis, D. A. and Lieberman, J. A. (2000). Catching Up on Schizophrenia: Natural History and Neurobiology [Review]. *Neuron*.

[15] Malaspina, D. et al. (n.d.). Advancing Paternal Age and the Risk of Schizophrenia.

[16] McGorry, P. D. et al. (2008). "Back to the Future," *Archives of General Psychiatry*, 65(1), p. 25.

[17] McGrath, J. et al. (2008). "Schizophrenia: a concise overview of incidence, prevalence, and mortality," *Epidemiologic Reviews*, 30, pp. 67–76.

[18] McLoughlin, B. C. et al. (2014). "Cannabis and schizophrenia," *The Cochrane Database of Systematic Reviews*, 2014(10), p. CD004837.

[19] Stilo, S. A. and Murray, R. M. (2010). "The epidemology of schizophrenia: replacing dogma with knowledge," *Dialogues in Clinical Neuroscience*, 12(3), pp. 305–315.

[20] Vaucher, J. et al. (2018). "Cannabis use and risk of schizophrenia: a Mendelian randomization study," *Molecular Psychiatry*, 23(5), pp. 1287–1292.

[21] Chr Kyziridis, T. (no date). "Notes on the history of schizophrenia," *German Journal of Psychiatry* [Preprint]. Available at: http://www.gjpsy.uni-goettingen.de.

[22] Shafiei, A. (2021). "Prediction modelling using logistic regression and constrained confidence partitioning: remission in real-world samples of patients living with schizophrenia," *Journal of Psychiatric Research* [Preprint].

[23] Orrù, G. et al. (2020). "Machine learning in psychiatry: A systematic review," *Neuroscience & Biobehavioral Reviews*, 36(1), pp. 58–70.