University of
St Andrews

Machine Learning: A Novel Approach for Pathology in Mast Cell Diseases

A Proof of Concept

Lea Brody-Heine | ID: 230008959
Co-Author: Lena Korfmacher
Supervisor: Professor Tom Kelsey

In partial fulfillment to obtain the degree
MSc Computing & IT

August 13, 2024

Abstract

In this dissertation, we applied machine learning techniques to improve the diagnosis and research of mast cell diseases, focusing on idiopathic mast cell activation syndrome (i-MCAS) and hereditary alpha-tryptasemia (HαT). We developed and integrated machine learning models that utilized both pathological tabular data and biopsy images to build a comprehensive proof-of-concept tool. To overcome the challenges of limited datasets, we compared the success of traditional data augmentation and generative models, followed by the training and testing of various machine learning algorithms. Despite identifying some limitations in generating accurate synthetic data, the traditional augmentation method proved more successful. To further explore i-MCAS, I created a tool to analyze microbiome data and investigate microbial discrepancies in i-MCAS. Our proof-of-concept tool demonstrates the potential for machine learning to discover novel patterns in complex pathological data, though, we acknowledge the limitations in collecting and integrating image and tabular data. Further refinement and integration into existing diagnostic frameworks is necessary for broader adoption.

Declaration

I hereby certify that this dissertation, which is approximately 14720 words in length, has been composed by me, that it is the record of work carried out by me, and that it has not been submitted in any previous application for a degree. This project was conducted by me at the University of St Andrews from 01/2024 to 08/2024 towards the fulfillment of the requirements of the University of St Andrews for the degree of MSc Computing & IT under the supervision of Professor Tom Kelsey.

For the protection of confidential data, parts of the data are omitted prior to submission to the Director of Postgraduate Teaching, our supervisor, and second grader.

Lea Brody-Heine
August 13, 2024

Table of Contents

List of Figures

List of Tables

1 Introduction

The primary goal of this dissertation is to explore the potential of machine learning (ML) methods in supporting clinical diagnosis and research for mast cell diseases (MCDs), rather than providing definitive scientific evidence. We apply various machine learning (ML) and deep learning (DL) models to both pathological tabular data and biopsy image data, demonstrating how these technologies can be integrated into a cohesive framework for medical research and diagnostics. In doing so, we developed a tool that clinicians can use to make discoveries and diagnoses.

The dissertation is structured as follows: I begin with a background section that provides context and a literature review. This is followed by a presentation of the methods and results, where I discuss the development, evaluation, and results. The discussion section considers challenges and limitations, broader implications, and future directions. Finally, the dissertation concludes with idealized recommendations for using the tools and reflections on personal contributions to the project.

1.1 Problem Statement & Motivation

Diagnosing MCDs presents varying levels of difficulty. Hereditary alpha-tryptasemia (HαT) and mastocytosis show clear genetic and blood abnormalities. However, diagnosing mast cell activation syndrome (MCAS) remains challenging due to the absence of definitive biomarkers. Diagnosis often relies on symptomatology and treatment response, which can lead to misdiagnosis or delayed treatment. This is particularly true for idiopathic MCAS (i-MCAS), where patients exhibit symptoms without clear pathological findings, complicating diagnosis and treatment [1].

The uncertainty surrounding i-MCAS highlights the need for novel diagnostic approaches, such as exploring microbiome discrepancies. Emerging evidence suggests the microbiome could influence immune responses and potentially trigger MCAS symptoms [2], yet studies integrating this aspect into the diagnostic framework are lacking. This dissertation aims to bridge this gap by utilizing ML techniques to analyze microbiome and other pathological data to find patterns that correlate with i-MCAS and other MCDs.

1.2 Research Objectives

Objective 1: Develop a ML Algorithm for Pathological Data Analysis

- Replicate findings from Hamilton et al. (2021) to ensure validity in pathological data analysis [3].
- Optimize the algorithm for precision and efficiency with large datasets.
- Validate the algorithm against traditional diagnostic methods.

Objective 2: Explore Generative AI Models for Synthetic Data Generation

- Implement Generative Adversarial Networks (GANs) to create high-quality synthetic tabular data.
- Apply traditional augmentation techniques to enhance data variability.

- Augment image data (see Lena Korfmacher's paper: *Exploring Mast Cell Diseases with Machine Learning, A Proof of Concept*).
- Test the effectiveness of synthetic data for training various ML algorithms.

Objective 3: Develop a Proof-of-Concept AI Tool for Clinical Application and Research

- Create a prototype integrating ML algorithms to address diagnostic and research challenges for MCDs.
- Demonstrate the tool's potential for aiding research into MCDs, particularly MCAS.

Objective 4: Build a tool for the Investigation of Microbiome Discrepancies in i-MCAS Patients

- Use data augmentation techniques to generate microbiome datasets for normal and i-MCAS patients. Manipulate specific bacteria classes to mimic a common discrepancy.
- Employ an unsupervised learning model to identify microbiome distinctions between i-MCAS patients and healthy individuals.

2 Context Survey

2.1 Machine Learning

2.1.1 Definition

ML is a branch of AI where computers learn and make decisions or predictions based on data [4]–[7]. Unlike traditional programming, where explicit instructions are given, ML models automatically learn patterns from data to perform tasks. Many modern ML applications have outperformed humans substantially on complex tasks [8]–[11].

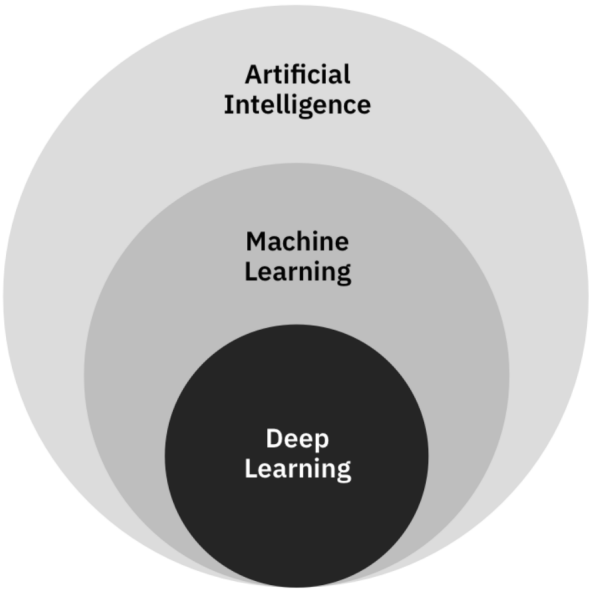
2.1.2 Application & Related Fields of Machine Learning

ML is widely used across various domains, including predictive analytics, intelligent decision-making, image and speech recognition, Natural Language Processing (NLP), and computer vision [12], [13]. In medicine, ML can support medical imaging, personalized medicine, predictive diagnostics, drug discovery, and pandemic forecasting [12]–[16]. In finance, it can aid algorithmic trading, fraud detection, and risk management [12]. ML also has applications in cybersecurity, recommendation systems, inventory management, customer service automation, autonomous vehicles, traffic prediction, smart cities, sustainable agriculture, and IoT [12], [13]. Additionally, the energy sector and Industry 4.0 utilize ML for smart manufacturing and quality control [13], [17], [18].

ML is a broad field focused on mimicking human intelligence and problem-solving abilities [12], [19]. Originating in the 1950s with pioneers like Turing, Samuel, and McCarthy, ML enables AI systems to learn and adapt [4], [20], [21]. Deep learning (DL), an advanced method of ML, has emerged due to increased data availability and computing power [12], [22]. DL, often seen as a subset of ML, uses neural networks

with multiple hidden layers for feature extraction and scalable learning, making it more effective for complex, unstructured data like text, speech, and images [19], [23].

Figure 1: *Machine Learning in the Context of Artificial Intelligence and Deep Learning*



Source: IBM

Computer vision is a subset of AI that enables computers to interpret visual inputs through DL, thus, equipping them to 'see' the world.[24]. ML also shares strong connections with statistics, as both fields focus on pattern recognition and prediction. While statistics emphasizes inference about populations, ML prioritizes generalizable predictions for forecasting future cases [6]. Both inference and prediction are valuable in research [6].

2.1.3 Learning Techniques

ML systems can be trained through various techniques. This dissertation focuses on supervised, unsupervised, and semi-supervised learning, which are commonly used for straightforward tasks. Reinforcement learning is more suitable for complex tasks that require consideration of the surrounding environment [13], [22].

Supervised Learning (SL) algorithms learn from a dataset containing both input variables and their corresponding labels [22]. The goal is to infer a function that maps inputs to outputs, encompassing tasks such as classification, which predicts categories, and regression, which predicts numeric values [13], [22].

Unsupervised Learning (UL) involves training algorithms on data without labels to identify patterns and group similar data points [22]. This is valuable for the exploration or discovery of structures and trends within data [13]. Common UL tasks include anomaly detection, association rule learning, clustering, and dimensionality reduction [13]. Dimensionality reduction decreases the feature space of a dataset, improving model performance and efficiency [22]. Anomaly detection identifies unusual values or outliers, while association rule learning infers relationships between attributes in large datasets [22].

Semi-supervised Learning (SSL) is a hybridization of SL and UL [13]. The algorithm is trained on a dataset that is partially labeled [22]. SSL is beneficial when labeled data is scarce and unlabeled data is numerous. This is common in many real-world scenarios as labeling data is usually time-consuming and costly [13], [22]. The motivation behind using SSL is to derive better predictions than an SL model on the few labeled instances alone [13].

ML systems can be further categorized by their learning process. Batch learning

processes data all at once and is typically time-consuming, whereas online learning updates models incrementally as new data arrives, making it suitable for systems that need to adapt quickly [22].

A model's performance relies on its ability to generalize to new data [12]. Generalization can be approached through instance-based learning, which uses a similarity measure derived from the training data, or model-based learning, which uses a model built from the training data to make predictions on new data [22].

2.1.4 Machine Learning Algorithms

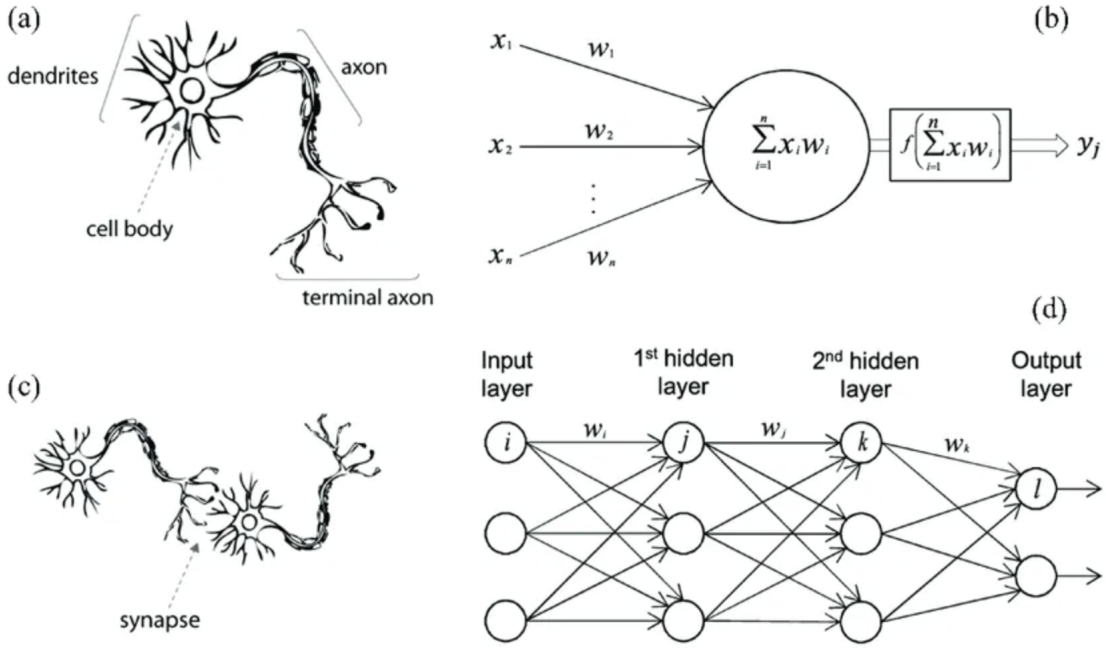
The ML algorithm landscape is broad, encompassing both traditional and modern approaches. Below is a selection describing algorithms relevant to this study:

- *Logistic Regression* is often used for binary classification but can also handle multiclass tasks using strategies like One-vs-Rest or Softmax Regression. It estimates the probability of an instance belonging to a class by calculating a weighted sum of input features and applying a logistic function, producing a probability between 0 and 1 [22].
- *Decision Trees* split input data into subsets through a series of decisions, creating a tree-like structure. The feature that best reduces impurity (measured by entropy or the Gini coefficient) is chosen at each node, with the leaf nodes representing outcomes. Decision Trees are intuitive and easy to interpret, making them a transparent, white-box model [22].
- *Random Forests* are ensemble methods that combine several Decision Trees trained on different data subsets using bootstrap sampling. The final prediction aggregates the predictions of individual trees, typically resulting in better performance than a single Decision Tree [22].
- *K-Means* [25] is a popular UL algorithm that groups n unlabeled instances into k clusters, where each instance belongs to the cluster with the nearest mean (centroid). The algorithm starts by selecting k initial centroids and assigning each instance to the nearest one. Centroids are updated based on the mean of the instances in each cluster, and the process repeats until centroids stabilize. The goal is to minimize within-cluster variance, though K-Means can be sensitive to initial centroid selection and the choice of k [26].
- *Principal Component Analysis (PCA)* is a dimensionality reduction technique that projects data onto a lower-dimensional subspace by finding the principal components, i.e., the directions in which the data varies the most. PCA reduces the number of features while retaining most of the variability in the data [22].
- *Boosting Algorithms* improve model robustness by combining predictions from several base estimators. Models are trained sequentially with each new model focusing on errors made by previous ones. Popular boosting methods include AdaBoost, Gradient Boosting, and XGBoost [22].
- The *Support Vector Machine (SVM)* is a classification algorithm that identifies the hyperplane that best separates the classes in the feature space. It maximizes the margin between the closest points of the classes (support vectors) [22].

Recent advancements in ML, driven by increased data availability and enhanced computational power, have led to the growing prominence of DL [12], [22]. Artificial Neural Networks (ANNs), particularly Multi-Layer Perceptrons (MLPs), now often outperform traditional ML algorithms on complex, large-scale problems [22]. ANNs are inspired by biological neurons and consist of an input layer, one or more hidden layers, and an output layer. Each input is weighted, summed, and passed through an activation function to produce an output, with backpropagation adjusting weights to minimize errors and improve performance [22], [27]–[29]. Figure 3 illustrates an exemplary architecture of a simple ANN.

Figure 2 provides a simplified illustration of both biological neurons and synapses compared to artificial neurons and ANNs.

Figure 2: *Biological and Artificial Neural Networks: (a) biological neuron; (b) artificial neuron; (c) biological synapse; (d) ANN synapses*

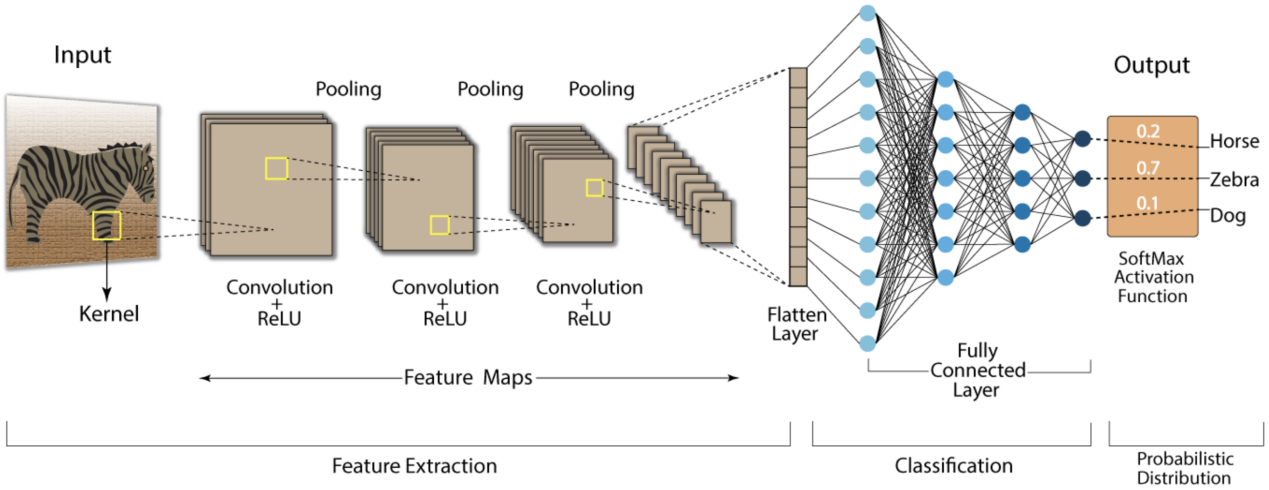


Source: [30] as cited in [31]

Generative Adversarial Networks (GANs) are another DL model that involves two neural networks–a generator and a discriminator. The generator network produces fake samples based on the training data, and the discriminator network examines the generated samples and predicts whether a sample is real or fake. This mechanism is based on game theory; both counterparties are incentivized to minimize their own cost. The generator tries to generate samples that will incorrectly be classified as real, while the discriminator aims at correctly detecting real and fake data [32]. GANs are one of the most powerful DL architectures and have been applied to numerous tasks. However, the advances in generative AI have enhanced deepfakes and misinformation, and concerns about the ethical use of AI have grown [12], [33].

Ensemble learning combines multiple algorithms to improve prediction accuracy by correcting individual errors [34]. Ensemble methods can be dependent, with each base learner influencing the next, or independent, meaning that the base learners' outputs do not impact the others [34]. Popular ensemble techniques include bagging, boosting, and stacking [22].

Figure 3: *Exemplary CNN sequence in a multiclass classification task*



Source: developersbreach.com (2020)

2.1.5 Common Challenges

An effective and efficient ML solution depends on the data and the learning algorithm [12], [13], [22]. The quantity, representativeness, and quality of data–in addition to the relevance of features–are essential to the model's accuracy and its ability to generalize well [12], [22]. However, real data is often scarce or incomplete, which can negatively impact the success of an ML algorithm [12], [13]. This has high relevance in a medical context, where labeling depends on expert knowledge and the collection of sufficient training data which is both expensive and difficult [35].

Determining the amount of training data required to achieve optimal accuracy for an ML model is complex and lacks a standard approach. Research generally indicates that the accuracy of predictions tends to improve as the volume of training data increases [22], [36]–[38]. This is particularly relevant in DL, where increasing the amount of data notably enhances performance. Whereas traditional ML algorithms exhibit only marginal improvements with additional data, underscoring the data-intensive nature of DL compared to conventional methods [23]. It was found that different amounts of training data can result in varying performance outcomes when comparing different algorithms [39]. Additionally, the required amount of training observations depends on the amount and complexity of features in a dataset. Datasets with simple patterns usually require fewer observations than datasets with complex patterns and relationships [40].

Although research has shown that increasing the amount of data is generally more effective than choosing or fine-tuning the right algorithm [36], it is important to consider that small or medium-sized datasets are common [22]. Standard techniques to handle limited data availability are: data augmentation, data simulation, and transfer learning [23]. Data augmentation and simulation increase training data by artificially augmenting existing data or generating new data. Data augmentation is particularly effective for image data through techniques like rotating, mirroring, cropping, and augmenting the color space (brightness, contrast, hue). Nevertheless, not all of these techniques are suitable for data when the augmentation incorrectly changes important characteristics of the existing data [23]. To verify that synthetic data preserves the statistical

properties of the original dataset, methods like the Kolmogorov-Smirnov test can be used to compare distributions [41].

Transfer learning involves either using data from similar tasks to improve the original input or employing a pre-trained model and fine-tuning the last layers with the available data. A DL network learns biases and weights during its initial training on a large dataset and transfers them to another network for retraining a new model. This is beneficial since the new model uses pre-trained weights instead of starting its training from scratch, which saves computational costs and time and enhances network generalization [23]. Many famous CNN models were trained on extensive datasets and have been successfully applied in various domains, including medical research [35], [42], [43].

Overfitting occurs when a model performs well on training data but fails to generalize to new data. It occurs when the complexity of the model is too high relative to the amount and noisiness of the training data, and thus, it will identify patterns in the noise [22]. Addressing overfitting involves collecting more data, reducing noise, or simplifying the model by regularization, feature reduction, or using a less complex model [22]. Underfitting, by contrast, happens when a model is too simple to capture the data's underlying structure. This can be corrected by using a more complex model, enhancing features, or reducing constraints like regularization [22]. Balancing model complexity to avoid both overfitting and underfitting is key to building robust models.

Overall, the literature shows that data is of key importance for effective ML solutions, but that it is often challenging to obtain sufficient good-quality data. Ultimately, both ML projects and their underlying data can vary greatly. Specific conclusions regarding the optimal amount of training data and the choice of the best-performing algorithm require consideration on a per-case basis.

2.1.6 Regulation and Explainability

AI's growing use has raised a number of concerns. Regulatory frameworks are becoming increasingly stringent, restricting the free handling of data and the use of non-transparent AI methods [44]. Two of the most influential legislations are the EU General Data Protection Regulation (GDPR) [45] and the EU Artificial Intelligence Act (AI Act) [46]. The GDPR sets precise requirements for data processing, controlling, and data subjects' privacy rights. The AI Act is the first legal framework to formulate requirements and obligations for AI systems following a risk-based approach.

Concerns regarding the responsible use of AI are particularly relevant in the medical field, due to the sensitive nature of the data and its impacts on human life [44]. Smith noticed that clinicians are legally required to justify their actions, whereas technologists only follow ethical codes of practice [47]. Further, she discovered there is no clear consensus on who is responsible for the outcomes of AI systems. While the literature agrees that users should evaluate the outputs of AI systems before applying them clinically, technologists remain only responsible for the accuracy of their systems but not for the clinical impacts. Finally, there is currently no case law to guide negligence and liability for AI systems [47].

At the center of these concerns is explainability. The term explainability is closely related and mostly used interchangeably with the term interpretability, although some

researchers have tried to distinguish them [48]. In our dissertation, interpretability is treated as synonymous with explainability [49]. While simpler models like Decision Trees are relatively easy to explain, tracing decisions made by complex models like ensembles or deep neural networks is challenging and often infeasible [22], [48]. These models, often referred to as “black boxes”, face criticism for their lack of transparency, particularly in fields like medicine, where the inability to fully justify AI-driven decisions can pose considerable safety risks [12], [44], [47].

Recent research has focused on improving the transparency of AI models by developing methods for explainable AI (xAI) [12]. The explainability of an ML model can be defined as “the degree to which an observer can understand the cause of a decision” [49]. The objective behind xAI is to provide explanations that make the system’s reasoning clear to users [50]. An xAI can therefore be defined as an AI system that 1) explains its reasoning to a human user, 2) specifies its strengths and weaknesses, and 3) provides an understanding of its future behavior [50].

The following taxonomy of xAI techniques identifies four categories of explainability methods [48]: Methods with the purpose of 1) creating white-box models that have intrinsic explainability, 2) explaining black-box models post-hoc, 3) enhancing fairness of a model, and 4) testing the sensitivity of predictions made by a model. Further, xAI methods can be distinguished by the scope of their explainability through their use of local or global methods and model-specific or agnostic methods [48]. Local methods provide an instance-specific explanation, whereas global methods aim to explain the entire model. Model-specific methods only apply to a particular model or a group of models, whereas agnostic methods apply to any model.

Various general-purpose xAI methods have been developed to visualize how features interact and contribute to the predictions of complex black-box models after they have been trained. But, building interpretable models–white-box models–that also achieve high performance is generally considered challenging [48]. In image classification, a common explainability approach is to generate saliency maps, which highlight the areas on an image that were most critical for the classification decision [48]. One of the most popular approaches is Gradient-weighted Class Activation Mapping (Grad-CAM), which captures the gradient information of a target class in the final convolutional layer to create a localization map that marks the relevant regions in the image for predicting the class [51].

In addition to individual explainability methods, several xAI evaluation tools have been developed, such as Quantus, which enables the comparison of different xAI methods using various metrics [52], and CLEVR-XAI, a benchmark dataset designed for evaluating the accuracy of neural network explanations against ground truth [53]. Despite these advancements, xAI is still considered to be in its early stages and lacks well-defined formal mathematical definitions and standardized metrics [48].

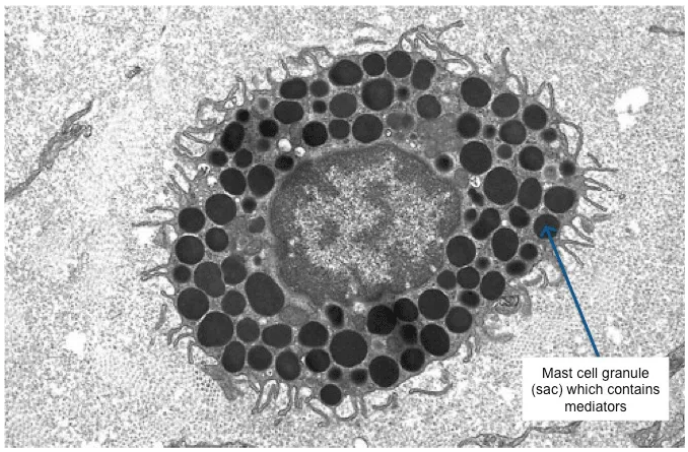
2.2 Mast Cells & Mast Cell Diseases

2.2.1 Mast Cells

Mast cells (MCs) are a type of white blood cell found in loose (areolar) connective tissue in nearly every part of the body. These immune cells are derived from the myeloid

lineage – or bone marrow [54]. Mature MCs can be found in the gastrointestinal tract, skin, and respiratory system, among other organs. They are also found near blood vessels, nerve cells, and lymph vessels. MCs play a pivotal role in keeping humans healthy by releasing inflammatory mediators that protect against parasites and infections [55].

Figure 4: *Mast Cell (electron micrograph)*



Source: Image provided by Mariana Castells, MD, PhD

MCs are mononuclear with secretory granules that release mediators during allergic or other MC reactions (see Figure 4) [55]. MCs can be triggered by: venoms, foods, drugs, infections, and in MC dysfunction: also friction, stress, or temperature changes [56]. IgE receptors, found in the plasma membrane of MCs, bind to circulating IgE. When IgE binds to MCs, degranulation occurs. This is the process of releasing inflammatory mediators into the extracellular space from the granules within the MCs [54].

MC mediators, such as histamine, leukotrienes, prostaglandins, tryptase, interleukins, heparin, and tumor necrosis factor-α, are selectively released during MC events, causing various symptoms and the regulation of bodily functions (see Figure 5 [56]. Histamine, for example, binds to H1 and H2 receptors throughout the body, affecting neurons, blood vessels, the gastrointestinal tract, and more. It regulates functions like stomach acid secretion, wakefulness, and hunger. However, high levels of histamine can lead to MC symptoms [57].

There are two types of MCs: MC(T) and MC(TC). MC(T)'s granules are predominantly filled with tryptase and are located near mucosal tissues exposed to external environments–such as the gastrointestinal and respiratory mucosa. The pro-inflammatory actions of MC(T)s lead to the recruitment of other circulating immune cells, increasing mucus production or enhancing gut peristalsis to minimize pathogen invasion [54].

Figure 5: *Symptoms of Mast Cell Mediators*

MEDIATOR	POSSIBLE EFFECTS
Histamine	Flushing, itching, diarrhea, hypotension
Leukotrienes	Shortness of breath
Prostaglandins	Flushing, bone pain, brain fog, cramping
Tryptase	Osteoporosis, skin lesions
Interleukins	Fatigue, weight loss, enlarged lymph nodes
Heparin	Osteoporosis, problems with clotting/bleeding
Tumor Necrosis Factor-α	Fatigue, headaches, body aches

Source: Carter MC, et al. Immunol Allergy Clin North Am. 2014;34(1):181-96.
Theoharides TC, et al. N Engl J Med. 2015;373(2):163-72.

MC(TC)'s granules contain tryptase, chymase, and carboxypeptidase and are found in the submucosa and connective tissues, adjacent to the skin, blood vessels, and lymph vessels. MC(TC)s are essential for tissue repair. Following an injury, they release inflammatory markers, independent of IgE mechanisms. Subsequently, the secretion of heparin, tryptase, and tissue plasminogen activator (t-PA) by these cells regulates blood flow to enhance nutrient and immune cell delivery [54].

MCs have cytoplasmic granules that cause metachromasia–a property observed during lab staining procedures–where specific dyes are used to visualize cells and tissues under a microscope. This technique is unique to MCs and basophils. For example, toluidine blue stains granules red-violet, but they turn pink once expelled [54]. MCs can be distinguished from basophils by their mononuclear morphology or special stains like tryptase or monoclonal antibodies targeting the MC kit (CD117) receptor. Hematoxylin and eosin staining, though revealing eosinophilic granules, does not reliably differentiate MCs from eosinophils. Under light microscopy, MCs appear as oval or irregularly shaped cells with dense granular cytoplasm often obscuring the nucleus [54].

The following describes the different functional states of MCs:

- *Intact Cells:* Found near blood vessels and within the deeper layers of the dermis and subcutaneous tissue. These cells have densely packed granules, making other cellular structures difficult to discern. They are spindle-shaped.
- *Spreading Cells:* These cells have many granules, though fewer than intact cells, and are located in the superficial connective tissue near the upper dermis.
- *Degranulated Cells:* These cells lose their metachromatic properties, appearing pale with a light pink stain and a blue nucleus (after lab staining).

MCs' strategic positioning highlights their crucial role in various physiological and pathological processes. There has never been a noted deficiency in MCs within medical history, underscoring their essential function for sustaining human life [54].

2.2.2 Mast Cell Diseases

MCDs can be from abnormal MC growth, accumulation, and/or improper mediator release, leading to symptoms across multiple organ systems [55]. Symptoms like flushing, itching, diarrhea, hypotension, and shortness of breath primarily result from MC degranulation [54], [56]. MC dysfunction can be localized or systemic and impact entire biologic systems [1].

There are three primary types of MCDs: mastocytosis, MCAS, and HαT. These conditions cause extensive discomfort and disability due to the release of MC mediators and/or the infiltration of MCs in critical organs. Although systemic mastocytosis is rare, the incidence of MCAS has been increasing [55]. A key phenomenon in MC pathology is type I immediate hypersensitivity reaction. IgE antibodies bind to antigens, circulate, and attach to Fc receptors on MCs. Upon antigen encounter, IgE receptor crosslinking triggers MC degranulation [54].

MCDs are classified into three main classes: primary, secondary, and idiopathic

Table 1: *Classification of MC Disorders*

MCD Classes	Clinical Entities	Diagnoses
Primary	Clonal Expansion of MCs	Mastocytosis (cutaneous and systemic), monoclonal mast cell activation syndrome (MMAS), mast cell sarcoma, mast cell leukemia, mastocytoma
Secondary	Underlying Pathological Process	IgE-mediated hypersensitivity reactions, urticaria, angioedema, food and drug anaphylaxis, mast cell hyperplasia (owing to systemic diseases such as chronic infection or autoimmune disease)
Idiopathic	Unknown Cause	Idiopathic Mast Cell Activation Syndrome (i-MCAS), idiopathic anaphylaxis, chronic idiopathic urticaria/angioedema

Source: Adapted from Khokhar & Akin, 2019 & Leru 2022 [1], [58]

(see Table 1). Primary MCDs include clonal disorders like mastocytosis, monoclonal mast cell activation syndrome (MMAS), mast cell sarcoma, mast cell leukemia, and mastocytoma, all involving clonal expansion of mast cells in various tissues. Secondary MCDs result from other pathological processes, such as IgE-mediated hypersensitivity, urticaria, angioedema, or mast cell hyperplasia due to chronic infections or autoimmune diseases. Idiopathic MCDs, such as i-MCAS, idiopathic anaphylaxis, and chronic idiopathic urticaria/angioedema, occur without a known cause [1].

Understanding MC pathology is essential for recognizing and managing MCDs. These diseases present a range of symptoms and can greatly impact a patient's quality of life [1], [55].

Mast Cell Activation Syndrome:
MCAS involves the abnormal release of mast cell mediators, affecting multiple organ systems. Although MCs in MCAS patients appear normal, they exhibit excessive activation, thus, releasing large amounts of inflammatory mediators [55], [59].

Patients with MCAS frequently experience recurring symptoms such as flushing, itching, hives, anaphylaxis, abdominal pain, diarrhea, and neurological issues like headaches, brain fog, and fatigue. Triggers include foods, drugs, environmental changes, physical stimuli, stress, infections, and even pain [55].

Diagnosing MCAS is complex, requiring the identification of elevated mast cell mediators during symptomatic episodes and ruling out other conditions with similar symptoms. Commonly measured mediators include serum tryptase, histamine, prostaglandin D2, and leukotriene E4. However, collecting samples with these mediators is challenging and often leads to false negatives. A positive response to medications that block or inhibit mast cell mediators is a key diagnostic approach for i-MCAS [59].

Treatment for MCAS focuses on avoiding triggers and using medications that block MC mediators. Common treatments include H1 and H2 antihistamines, leukotriene inhibitors, mast cell stabilizers like cromolyn sodium, and aspirin to reduce prostaglandin production. In severe cases, anti-IgE therapy or other immunomodulatory treatments may be used [59]. However, these treatments do not cure or put i-MCAS into remission; they only manage symptoms.

Hereditary Alpha-Tryptasemia (HαT):
HαT is a genetic trait found in approximately 5% of people in Western Europe and the United States. It results from extra inherited copies of the TPSAB1 gene, leading to elevated baseline serum tryptase (BST) levels–typically above 8 ng/mL. This is due to increased production of α-tryptase, a precursor form of the enzyme tryptase released by mast cells [60].

HαT is inherited in an autosomal dominant manner, meaning one copy of the mutated gene from a parent is sufficient to cause the trait. While elevated tryptase levels do not have a known direct impact on health, they are often linked to symptoms related to mast cell activation [60].

Clinical manifestations of HαT can vary widely. Some individuals may experience severe symptoms while others may be asymptomatic or have mild symptoms [60]. Diagnosis involves measuring serum tryptase levels, and if elevated, genetic testing is conducted to determine the number of TPSAB1 gene copies. This is important for distinguishing HαT from other mast cell disorders and understanding the cause of elevated BST levels [60].

Management focuses on treating symptoms and preventing severe reactions. Antihistamines, leukotriene inhibitors, MC stabilizers, and monoclonal antibodies like omalizumab are used to manage symptoms. Avoidance of known triggers and regular monitoring are also key to patient's health [60].

Mastocytosis:
Mastocytosis is a group of disorders characterized by the abnormal accumulation of MCs in multiple organs. Pathogenesis typically involves mutations in the KIT gene, promoting uncontrolled MC growth [61], [62]. There are two primary forms: cutaneous mastocytosis (CM) and systemic mastocytosis (SM).

Cutaneous Mastocytosis mostly affects children and presents with skin lesions, which may indicate deeper systemic involvement in adults. The primary subtypes include:

- Maculopapular Cutaneous Mastocytosis (MPCM) or urticaria pigmentosa, characterized by red-brown skin lesions.
- Diffuse Cutaneous Mastocytosis (DCM), noted for extensive skin thickening and blistering.
- Cutaneous Mastocytoma, often appearing as solitary skin nodules in children.

Systemic Mastocytosis ranges from indolent forms with good prognosis to aggressive types that substantially impact organ function and require complex treatments:

- Indolent Systemic Mastocytosis (ISM) is the mildest, often with symptoms from

MC mediators like flushing and gastrointestinal discomfort.

- Smoldering Systemic Mastocytosis (SSM) and Aggressive Systemic Mastocytosis (ASM) present more severe symptoms and increased MC burden.
- Systemic Mastocytosis with Associated Hematologic Neoplasm (SM-AHN) and Mast Cell Leukemia (MCL) are severe, often coinciding with other hematologic conditions.

Diagnosing mastocytosis involves clinical evaluation, skin biopsy for CM, and bone marrow biopsy for SM. Key diagnostic criteria include the presence of multifocal dense infiltrates of MCs in the bone marrow or other extracutaneous organs, detection of KIT mutations, abnormal MC morphology, and elevated serum tryptase levels [62]–[64].

The treatment of mastocytosis focuses on managing symptoms and preventing complications. For CM, topical steroids and antihistamines may alleviate skin symptoms. For SM, treatment includes antihistamines, leukotriene inhibitors, MC stabilizers, and in severe cases, cytoreductive therapies. Patients with severe allergic reactions may require epinephrine for anaphylaxis. Early diagnosis and tailored therapeutic strategies are crucial for managing the disease and improving patient outcomes [61], [62].

2.3 Machine Learning & AI in Medicine

AI and ML are transforming the biotechnology industry by improving diagnostics, accelerating therapeutic development, and enhancing personalized medicine. These technologies can analyze complex biological data at unprecedented speeds and accuracy to predict disease progression and optimize treatment protocols. For instance, Google DeepMind's AlphaFold 3 has significantly advanced protein structure prediction, facilitating a more in-depth understanding of diseases and aiding in drug discovery [65].

2.3.1 Diagnostics and Doctor Assistance

Supervised ML models have greatly improved diagnostic accuracy in medical imaging. They are remarkably effective in early disease detection, such as predicting dementia from brain MRIs well before clinical symptoms appear [66]–[69]. Unsupervised learning techniques like clustering and principal component analysis have revealed latent patterns in data. This benefits patient stratification and personal therapy customization [67]. AI tools in clinical systems support treatment decisions and risk assessments by providing predictive analytics based on both longitudinal and real-time data [70].

Despite ML's many advancements, challenges regarding data heterogeneity, model interpretability, and the need for large annotated datasets to train strong models remain [71]. The integration of ML in biotechnology not only poses technical challenges but also raises ethical concerns regarding data privacy, model bias, and the transparency of AI decisions. Addressing these issues is essential to gain public trust and facilitate wider adoption of these technologies [72]. Ensuring that AI systems are transparent and interpretable is crucial for their acceptance in clinical practice, as clinicians need to understand and trust the decisions made by these systems.

2.3.2 AI & ML in Knowledge Discovery for Medical Research

AI and ML play key roles in medical research by identifying new biomarkers and understanding disease mechanisms, which improves therapeutic strategies. For instance, ML has identified biomarkers for diseases like interstitial cystitis and psoriasis, supporting the development of targeted therapies [73], [74].

In microbiome research, ML models have identified specific microbial signatures linked to specific diseases. In doing so, researchers can develop targeted therapies [75]. AI advances in genomics have also improved immune cell categorization and disease progression analysis, which are also critical for developing effective treatments [76].

Unsupervised ML techniques, such as clustering, enable the discovery of disease-related patterns without predefined labels, leading to novel insights [77]. These methods are essential in identifying patient subgroups and refining precision medicine approaches, where treatments are tailored based on individual genetic and molecular profiles.

Deploying AI and ML in medical research presents challenges as well. Concerns such as data privacy, algorithmic bias, and the need for powerful computational resources must be addressed. Ensuring the transparency and interpretability of AI models is crucial for their acceptance in clinical practice [78]. Establishing an ethical framework is essential to guarantee models promote fairness and equity.

Looking ahead, AI and ML in medical research is expected to grow substantially with advancements in computational power and algorithm design. Interdisciplinary collaborations and model refinements will likely lead to breakthroughs in medical science, enhancing our understanding of diseases and leading to innovative diagnostic and therapeutic tools [79].

AI and ML are reshaping medical research by enabling the exploration of complex biological data. As these technologies evolve, they promise to deepen our understanding of health conditions, refine diagnostic processes, and revolutionize treatments.

2.3.3 AI & ML in Drug Discovery

AI and ML technologies have greatly improved the drug discovery process. These technologies increase the success rates of new therapeutics. ML models can optimize and evaluate therapeutic candidates, potentially shortening development timelines and reducing costs.

ML models, predominantly deep learning, are effective in predicting drug-target interactions, optimizing drug candidates, and identifying potential toxicities early in drug development [80], [81]. By analyzing large datasets, ML models reveal patterns and relationships that conventional methods may miss to help researchers make more informed decisions.

These models can improve target identification by integrating genomic, proteomic, or chemical data to predict and validate new drug targets. This approach refines our understanding of disease mechanisms and potential intervention points [82]. Researchers are able to develop drugs with higher chances of effectiveness by identifying promising targets.

A noteable example is DeepMind's AlphaFold, which revolutionized protein structure prediction. The AlphaFold model allows researchers to discover molecular mechanisms of diseases and identify potential drug targets. This, in turn, accelerates the early stages of drug discovery [65].

In microbiome research, these technologies can be used to predict how drugs interact with the human microbiome, affecting drug efficacy and safety. By analyzing microbiome data, ML models can identify potential adverse effects early in the drug development process, making the development of therapeutics safer and more effective [75]. Furthermore, ML and AI technologies can develop microbiome-targeted therapies like probiotics and prebiotics. For example, ML algorithms can identify microbial strains that treat or prevent diseases linked to dysbiosis by analyzing microbiome data [83].

During the hit-to-lead conversion phase, ML algorithms optimize lead compounds by predicting pharmacokinetic and pharmacodynamic properties. This accelerates drug development and improves the likelihood of clinical trial success by prioritizing candidates with higher efficacy and safety potential [84]. Predictive toxicology models further enhance this process by forecasting adverse effects early, reducing the risk of costly late-stage failures [85].

When integrating ML into drug discovery workflows current challenges include data quality, model interpretability, and the need for substantial computational resources. Addressing these challenges requires ongoing algorithmic advancements, better data-sharing practices, and interdisciplinary collaboration [86]. Ensuring ML models are trained on high-quality data and their predictions are interpretable is essential for successful integration into drug discovery.

3 Method and Results

3.1 Programming Environment and Frameworks

We used Python [87] as the primary programming language for the development and implementation of our models and algorithms due to its simplicity and extensive library support. The majority of our project was conducted in Jupyter Notebooks [88] as they provide a suitable environment to add headlines and explanatory annotations, and display outputs and visualizations. We used Python scripts to import custom utility functions, where applicable. To see a full outline of our process refer to Appendix I.

The following ML and DL frameworks were used:

- *SciKit Learn [89]:* A Python library that provides simple tools for data analysis and ML. In this dissertation, SciKit Learn was utilized for preliminary data exploration and processing, feature engineering, and the implementation of traditional ML algorithms.
- *TensorFlow/Keras:* TensorFlow [90] is a large-scale, open-source ML system. Keras [91] is a high-level API for building and training DL models, running on top of TensorFlow. We used TensorFlow/Keras for building and training deep neural networks for image classification.

- *PyTorch [92]:* An open-source DL framework known for its dynamic computation graph and ease of use. PyTorch was employed for various deep learning tasks, including developing and training neural networks for image and tabular data.
- *Bootstrapping:* I used bootstrapping to create multiple resampled datasets from the original data to estimate the distribution of a statistic and improve the robustness of our models [93].
- *GANs:* I used Generative Adversarial Networks (GANs) to generate synthetic tabular data. This technique enhances the training dataset by providing additional, varied samples, which helps improve the performance of our classification models [94].

3.2 Data

3.2.1 Tabular Data & Exploratory Data Analysis (EDA)

This dissertation uses data collected by Hamilton et al. [3] from patients at the Mastocytosis Center at Brigham and Women's Hospital, the University of Florida's Gastroenterology Division, and the NIH's Laboratory of Allergic Diseases. The dataset was provided as an Excel workbook with two sheets.

We initially inspected the workbook and found deviations from a normalized tabular structure, essential for most ML algorithms. To maintain data integrity, we merged the two sheets and then created a CSV file for further processing.

We first examined some general characteristics of the dataset and made certain adjustments before we split it into training and test sets. For many patients, multiple rows existed due to different observations, where only the first row was fully populated, and subsequent rows had blanks except for deviating values. This structure made a stratified shuffle split impractical, as it could compromise data integrity and lead to bias in imputations. The order of the rows was critical in maintaining the data's integrity. Although splitting data early is best practice to ensure unseen instances in the test set, we delayed this step to maintain data quality, acknowledging this as a suboptimal approach for testing.

The dataset included 73 observations from 52 subjects and 21 features (16 categorical, 5 numerical) after the initial cleaning. This involved removing empty and duplicate columns. The label, "Diagnosis," is multi-class, with approximately 50% of observations labeled "MCAS," and 25% each labeled "normal" and "HaT." We then split the data into an 80% training set (58 observations) and a 20% test set (15 observations), using a stratified shuffle split to maintain class ratios.

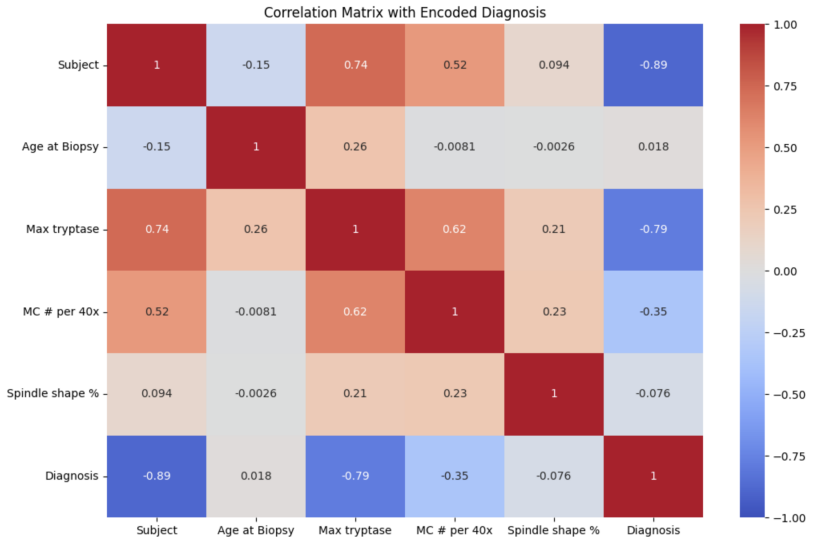
Key insights from the exploratory data analysis (EDA) include:

- The label classes were imbalanced, with "MCAS" being the most frequent (see Figure 15 in Appendix E).
- None of the numerical features followed a normal distribution; for example, *Max tryptase* was skewed towards lower values (see Figure 20 in Appendix E).
- All categorical features were imbalanced, with varying degrees across features.

- *Max tryptase* values were missing for ”normal” subjects, requiring scientifically informed imputation.
- Higher *Max tryptase* levels in HαT subjects and differences in MC # per 40x were consistent with previous studies [3]. The skews were as expected (see Figures Appendix E).
- Several outliers were identified in features such as *Spindle shape %* and *Max tryptase*, which could impact model performance and require further investigation (see Figure 16 in Appendix E).

We observed numerous distinct values within the categorical features, particularly *MC aggregate size and number*, *Procedure*, *Diagnosis.1 Assoc diagnoses*, *Labs (histamine or PGD)*, and *Medication*. These columns contained unstructured data with occasional typos and inconsistent naming, indicating the need for thorough feature engineering.

Figure 6: *Correlation: Features & Labels*



Source: Lena Korfmacher, *Small_Dataset_Model.ipynb*

The most significant correlations include *Diagnosis* and *Max tryptase* (-0.79), *Max tryptase* and *MC # per 40x* (0.62), and *MC # per 40x* and *Diagnosis* (-0.35) (see Figure 6), suggesting these features' importance.

As part of the EDA, I performed a clustering analysis with K-Means on numerical features using the first two principal components. I included the *Diagnosis* column as the label in the scatterplot. Then, I chose 3 clusters, assuming that the clusters would correspond to the 3 diagnosis classes. Interestingly, the resulting visualization shows that this was not the case. We noticed that different labels were present in each cluster and the classes were not entirely randomly spread across all clusters.

Finally, PCA on numerical features revealed that the first two components explained about 72% of the variance, with all four components necessary to explain 95% of the variance, underscoring the importance of all numerical features.

3.3 Data Cleaning & Preprocessing

3.3.1 Tabular Data Augmentation

With a total of 73 observations, the tabular data is limited for training a powerful ML model. While synthetic data cannot fully replace real data, it is beneficial for enhancing

model generalizability and robustness by generating additional samples [95]. However, the synthetic data must differ sufficiently from the original data, to be effective, as overly similar data would not add value [95]. GANs, commonly used for image data, are one option for tabular data augmentation, but they require converting categorical and discrete values into continuous ones, which can produce invalid data [95].

To address the scarcity of data, we applied two augmentation methods: First, we employed a traditional approach by copying and introducing noise. Then we attempted to create synthetic data using a GAN. Even though the augmentation typically focuses on training data [96], we also augmented the test data due to the small size of the test set, which had only 15 observations.

Traditional Augmentation:
We applied a conventional approach to data augmentation alongside a generative method to compare their efficacy. The literature on tabular data augmentation is limited, and optimal noise levels can vary, so we developed the following method based on practical ML community insights [96], [97].

1. *Data Expansion:*
   To address the limited size of the dataset, we first expanded the data by creating multiple copies of the original dataset. This was our foundation for later applying noise to generate varied instances for training.
2. *Introducing Noise:*
   Next, we introduced Gaussian noise with a mean of 0 and varying standard deviations to the numerical features. For categorical features, a percentage of values were randomly replaced with other unique values from the same column. The noise level, ranging from 5% to 1%, was applied and gradually decreased to identify the most effective level: 2%.
3. *Distribution Analysis:*
   We compared the resulting distributions with violin plots for numerical features and bar charts for categorical features. Our findings showed minor distribution differences between the original and augmented data. Categorical columns exhibited slightly more balanced classes post-augmentation, while numerical columns retained the same overall shape with more defined distributions.
4. *Augmentation Strategies:*
   We tested two augmentation strategies: a single-batch approach and an incremental three-batch approach. Both produced similar distributions, but the multi-batch method yielded greater variability, measured by the proportion of unique rows. Due to the risk of altering the original dataset's characteristics with the multi-batch strategy, we chose the single-batch method.

Results:
The traditional augmentation method generated approximately 20,000 training instances and 4,000 test instances. The noise levels were fine-tuned to ensure the augmented data closely resembled the original dataset. Despite the limited size of the original dataset, this approach provided a reliable augmentation method that maintained the integrity of the original data patterns.

GAN-based Data Augmentation:

1. *Data Preparation:*
   The cleaned medical dataset, including features like age, gender, diagnosis, and lab results, was loaded from a CSV file and inspected to verify suitability for GAN training and analysis. Numeric columns were scaled to the range (-1, 1) using MinMaxScaler, appropriate for the generator's tanh activation function. Categorical columns were encoded using OneHotEncoder. The processed features were combined into a single DataFrame.
2. *GAN Architecture:*
   The generator model, which took a latent vector (random noise) as input, was designed to generate synthetic data. It included layers such as Dense, LeakyReLU, and BatchNormalization to stabilize training and improve performance, producing both numerical and categorical data. The discriminator model, which evaluated real or generated data, included Dense, LeakyReLU, and Dropout layers to prevent overfitting and enhance the ability to distinguish real from synthetic data.
3. *Training the GAN:*
   The generator and discriminator were trained together in a loop. The discriminator was trained to differentiate between real and fake data, while the generator was trained to produce data that could fool the discriminator. Loss values for both models were monitored throughout the training process to assess performance.
4. *Synthetic Data Validation:*
   Initial checks on the synthetic data included visualizations and statistical tests, such as the Kolmogorov-Smirnov test [41], to compare the structure and distributions of synthetic data with the real data (see Figure 36 in Appendix K). This validation was crucial to affirm the synthetic data accurately represented real data distributions.

Results:
Initial checks on the synthetic data showed that while the GAN generated data, the distributions of most features did not closely match the real data. Histograms comparing the real and synthetic data revealed discrepancies (see Figure 32 in Appendix G). Therefore, indicating the synthetic data did not accurately represent the real data distribution. The Kolmogorov-Smirnov test [41] further confirmed these discrepancies, showing notable differences between the real and synthetic data distributions.

This discrepancy suggested that the model requires further fine-tuning and optimization. Due to the limitations of time for this project, traditional data augmentation methods were selected. These traditional methods provided more reliable results, ensuring that the augmented data closely matched the patterns observed in the real data.

3.3.2 Feature Engineering & Customized Preprocessing Pipeline

Based on the data exploration insights, we considered necessary imputations and transformations for all features. In most cases, this already included a numerical encoding (binary or ordinal) for categorical features. The amount of preprocessing for

this dataset was generally high as the dataset was manually created and originally not intended for an ML project.

- *Max tryptase:* Values were present for subjects diagnosed with MCAS and HaT but missing for control group subjects (i.e., diagnosed ”normal”). Our approach was to first forward-fill missing values and second impute new values for all normal subjects, based on a distribution with a median of 4.15 and a standard deviation of 2.12 in a range between 0 and 11.4, which was found to be a typical range of tryptase values [98].
- *Tryptase gene copy:* As outlined in the literature review, excess copies of the tryptase gene (TPSAB1) are characteristic of people diagnosed with HaT. More than 50% of the values in this column were missing, however, based on the paper by Hamilton et al. [3], we inferred that missing values indicated no excess tryptase gene copies. Further, it was evident from the study that none of the patients had more than one excess copy. Therefore, and given the small number of instances available, we refrained from a further differentiation between various alpha-beta constellations. We binary-encoded the column, setting missing values and ”don't have” to 0, and values indicating that one excess tryptase gene copy was present to 1.
- *Lymph aggregates:* We assumed that missing values indicated the absence of lymph aggregates and performed binary encoding.
- *Spindle shape %:* Missing values were present for instances with MC shape ”round”. Based on that, we assumed that missing values indicate a spindle shape of 0%, which we imputed accordingly.
- *MC aggregate size and number:* We assumed that missing values indicate no aggregates (clusters of MCs) and imputed 0 accordingly. In addition, we split up the column to reflect 1) the number of cells per aggregate and 2) the number of aggregates.
- *Procedure:* The column contained more unstructured values. We split the column into 1.) a column specifying the procedure (EGD or CLS), and 2.) a column specifying if there were any findings. Both columns were binary encoded.
- *Diagnosis:* Note that the dataset contained two columns named ”Diagnosis” prior to processing. The first diagnosis column was chosen as the label (containing the three classes normal, HαT, and MCAS); the second contained additional diagnostic details for some subjects. and associated diagnoses. Based on our insights from EDA, we first extracted all atomic values out of the text strings in both columns, which revealed that 89 distinct values were present. Second, we grouped these values into related medical condition groups to reduce dimensionality. This resulted in 8 new columns and we dropped the original columns.
- *Labs (histamine or PGD):* We dropped tryptase-related information as it was already included in the *Max tryptase* column. In addition, we split the column into two, one indicating whether or not elevated levels of histamine were present, the other one indicating whether or not elevated PGD levels were present (both binary encoded).
- *Medication:* Based on our insights from EDA, H1 and H2 medication was predominant. We transformed the medication column into three new columns, the

first of which indicates the presence of H1 medication, the second the presence of H2 medication, and the third the presence of other medication, each binary encoded. Patients who did not take medication therefore had a value of 0 in all three columns.

- *Villi/superficial, Lamina propria, Muscularis mucosae, Submucosa:* Values in these columns included ”x”, ”few”, and missing values, which we interpreted as an approximation for the number of MCs present in each of these tissue parts. We encoded these values with 1, 0.5, and 0.
- *CD63:* We applied binary encoding to indicate positive (1) and negative (0) staining of CD63. We based our imputation strategy on the percentages in the Hamilton et al. paper [3], with positive CD63 staining of 2/4 (50%) for HαT, 6/15 (40%) for MCAS, and 3/12 (25%) for control group patients.
- *Location of MC clusters:* Our transformation and imputing approach included splitting the column into four new columns with each binary value. We did not differentiate between ”bottom LP” and ”LP”. Further, we assumed that ”sm” is an abbreviation for submucosa. We note that the addition ”with lymph aggregate” has a discrepancy with the existing column ”Lymph aggregates”. We created a separate column ”Location of MC clusters - with lymph aggregates” to account for this discrepancy.

I bundled the feature engineering and preprocessing steps outlined above and applied them through custom functions in a customized pipeline (utils_ApplyPipeline.py). It resulted in 34 final features. I later applied the same pipeline to the augmented dataset to maintain consistent preprocessing (see Appendix H).

3.4 Tabular Data Models

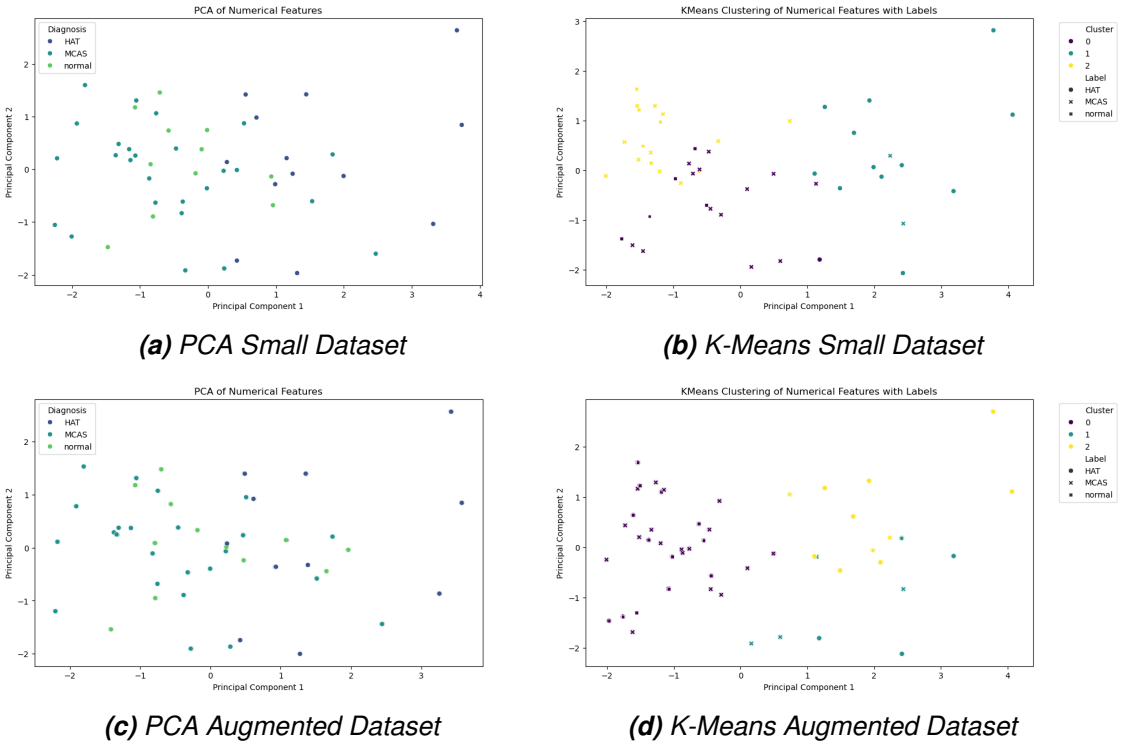
3.4.1 Data Exploration:

My customized script for data exploration analyzed both the original dataset and the augmented dataset. I used functions to visualize diagnosis counts, analyze distributions, detect outliers, and perform clustering and PCA. These standard functions made it easier to see the differences and similarities between the datasets. For example, the original dataset showed imbalances and many missing values, while the augmented dataset had similar patterns but with added variability to improve model training.

Using a standardized script for data exploration guarantees uniform analysis across datasets. This approach helps identify patterns and anomalies effectively, which might be missed with ad-hoc methods. The script simplifies the preprocessing workflow by covering various aspects of data exploration, from basic visualizations to complex analyses like PCA. This consistent process aids in comparing datasets, validating the impact of augmentation and making informed decisions for model training and evaluation. This methodical approach is essential for maintaining data integrity and maximizing the utility of both original and augmented datasets in medical research.

The original small dataset visualizations showed significant imbalances, such as a higher number of female patients and missing values for key features like tryptase levels. The distribution analyses highlighted specific peaks in age and max tryptase

Figure 7: *PCA and K-Means EDA Analysis*



(a) *PCA Small Dataset*

(b) *K-Means Small Dataset*

(c) *PCA Augmented Dataset*

(d) *K-Means Augmented Dataset*

Source: Lea Brody-Heine, *Traditional Aug Model Training.ipynb & Small_DataSet_Exploration.ipynb*

levels, indicating clustering around certain values. The PCA plot shows that while different diagnosis categories (HαT, MCAS, and normal) display some separation, there is noticeable overlap, particularly with MCAS. Similarly, the K-Means clustering plot reveals that clusters still contain a mix of diagnoses, especially cluster 0, which includes patients from all three categories (see Figure 7). This indicates that numerical features alone are not entirely sufficient to clearly distinguish all diagnoses, particularly MCAS.

The augmented dataset visualizations and analyses reflected more variability, which was expected due to the data augmentation process. The distribution of features remained consistent with the original dataset but showed a wider range and less pronounced peaks, indicating the addition of noise and synthetic diversity. Clustering and PCA analyses on the augmented data did not show clearer separations between diagnoses, particularly for MCAS, compared to the original data (see Figure 7). This suggests that the added variability from the augmentation did not greatly improve the model's ability to distinguish between conditions. The overlap between diagnoses remains similar in both datasets.

The comparisons show that while the original small dataset had specific patterns, the augmented data introduced more variability. Despite this, the added 2% noise did not improve the model's ability to distinguish between diagnoses, as the overlap between conditions remained similar. Therefore, the augmentation did not enhance the model's performance as much as expected.

3.4.2 Training on the Original Dataset:

The task is to predict the diagnosis (HαT, MCAS, normal) in a supervised, multiclass classification problem. Since the literature indicates that different algorithms perform differently in different use cases [39], we trained several different models suitable for this task:

- Logistic Regression (One-vs-Rest)
- Logistic Regression (Multinomial)
- Decision Tree
- Random Forest
- Gradient Boosting (GB)
- Support Vector Machines (SVM)
- Multilayer Perceptron (MLP)
- XGBoost

We performed an extensive grid search with cross-validation for each classifier to explore the best constellation of hyperparameters. As the explainability of AI outputs is highly relevant for clinicians, we included feature importance in the evaluation of the models to make results more explainable. We wanted to visualize which features contributed most to a classification decision.

However, not all models have a feature importance attribute, such as the Logistic Regression. Thus, feature importance was not applied. The results indicate that both the number of features contributing most to the classification decision as well as the importance of individual features vary greatly across different models. However, the features *Max tryptase* and *Autoimmune and Connective Tissue Disorders* were among the most important features across all the models (see Appendix F).

Three model training and evaluation rounds were performed. In the first round, models were trained and fine-tuned on the imbalanced dataset and the full feature space. In the second round, synthetic minority oversampling (SMOTE) was applied to balance the label classes [99]. In the final round, the feature space was reduced to the features that were most important across all models.

Decision Tree, Random Forest, GB, and XGBoost had the best performance through all three rounds. The best-performing models from the first round performed equally well on the balanced dataset, while the Logistic Regression (OvR), SVM, and MLP performed worse on the balanced dataset. Reducing the feature space did not have an impact on overall model performance for either of the models, indicating that dimensionality reduction could be useful to save computational cost and time. The best-performing models in all three training rounds achieved an accuracy of 93.33% with 95.00% precision, 93.33% recall, and 93.59% f1 Score (see Figure 8).

3.4.3 Training on the Augmented Dataset:

This section details the methodology and results of developing and training models on a tabular dataset to classify diagnoses related to MCDs.

Figure 8: *Testing results for model performance*

	accuracy	precision	recall	f1-score
Decision Tree	0.933333	0.95	0.933333	0.935873
Random Forest	0.933333	0.95	0.933333	0.935873
Gradient Boosting	0.933333	0.95	0.933333	0.935873
XGBoost	0.933333	0.95	0.933333	0.935873
Logistic Regression (One-vs-Rest)	0.866667	0.896667	0.866667	0.865608
Logistic Regression (Multinomial)	0.866667	0.896667	0.866667	0.865608
SVM	0.8	0.807143	0.8	0.798095
Neural Network	0.733333	0.764444	0.733333	0.730952

(a) *Performance overview with class imbalance and all features*

	accuracy	precision	recall	f1-score
Decision Tree	0.933333	0.95	0.933333	0.935873
Random Forest	0.933333	0.95	0.933333	0.935873
Gradient Boosting	0.933333	0.95	0.933333	0.935873
XGBoost	0.933333	0.95	0.933333	0.935873
Logistic Regression (Multinomial)	0.866667	0.896667	0.866667	0.865608
Logistic Regression (One-vs-Rest)	0.733333	0.754444	0.733333	0.730159
SVM	0.733333	0.754444	0.733333	0.730159
Neural Network	0.533333	0.546667	0.533333	0.514672

(b) *Performance overview with SMOTE and feature reduction*

Methodology:

1. *Preprocessing:*
   Initial inspections involved understanding the structure and distribution of the data, identifying missing values, and analyzing the class distribution of the target variable 'Diagnosis'. The training dataset included 20,416 samples with features such as 'Age at Biopsy', 'Max tryptase', 'MC # per 40x', and various categorical attributes like 'Gender' and 'Mast cell shape'.

   Both the training and test datasets were transformed using the customized pipeline script from utils_ApplyPipeline. This standardized the features post one hot encoding, specifically for the original associated diagnosis column. For a more exact comparison of the models trained on the original data and augmented data, we applied a standardized pipeline to this augmented dataset.

   Missing values were present in features like 'Lymph aggregates', 'Tryptase gene copy', and 'Location of MC clusters'. Missing values and empty strings were turned into NaN values. Additionally, all numeric features were transformed into numeric types. The label and features were then separated and a label encoder was applied to the y_train and y_test set. Numerical features were standardized using StandardScaler, and categorical features were encoded using OneHotEncoder to confirm all features were in a suitable format for ML models.

2. *Model Training:*
   We did not use every model trained on the small dataset here due to high computational expense. We chose models that utilized the extra GPU available to us via the lab PCs to make sure training and hyperparameter tuning were possible in a reasonable amount of time.

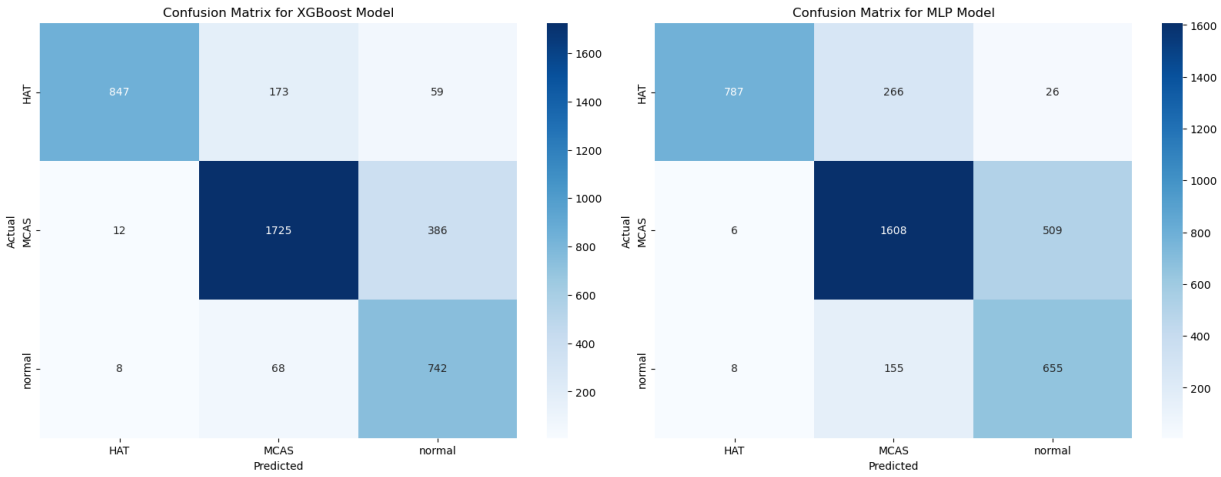
   An XGBoost classifier was trained on the preprocessed data. The model was

evaluated using accuracy, precision, recall, and f1-score metrics. XGBoost was chosen for its high performance and efficiency, particularly suited for handling large datasets and complex classification tasks. An MLP classifier was also trained and evaluated. This choice was driven by previous observations where the MLP performed poorly on a smaller dataset. We aimed to investigate whether the increased data size, augmented through traditional techniques, could improve its performance.

3. *Hyperparameter Optimization:*
   Hyperparameter tuning was performed using GridSearchCV to find the optimal parameters for the XGBoost model as it performed better than the MLP model. Parameters such as the number of estimators, max depth, learning rate, subsample ratio, and column sample ratio were explored to improve model performance.

Results:
The XGBoost classifier initially achieved an accuracy of 82.24%, with high precision and recall for HαT and MCAS, but lower precision for the normal class, leading to more false positives in that category (see Figure 9). These results suggest that while the model effectively identifies HαT and MCAS cases, it struggles more with accurately classifying normal cases.

Figure 9: *Confusion Matrix: XGBoost & MLP*



Source: Lea Brody-Heine, *Traditional Aug Model Training.ipynb*

The MLP classifier achieved an accuracy of 75.87%, showing a slight improvement over the small dataset. It exhibited balanced performance for MCAS but struggled with HαT and normal classes, resulting in more false negatives for HαT and false positives for normal cases (see Figure 9).

After hyperparameter optimization, the XGBoost model's accuracy improved to 91.17%, with strong performance across all classes. The best parameters included 'colsample_bytree': 0.6, 'learning_rate': 0.1, 'max_depth': 15, 'n_estimators': 200, and 'subsample': 0.6. This optimization enhanced the model's ability to generalize to unseen data, indicating the importance of fine-tuning for improving model performance.

However, applying SMOTE and SMOTEENN techniques [99] to address class imbalances led to a drastic reduction in model performance, with accuracy dropping to around 20%. This suggests that these techniques introduced noise and artificial patterns that the models could not effectively learn from. This highlights the need for alternative strategies to handle class imbalances in medical datasets.

The augmented dataset yielded mixed results. The XGBoost model showed considerable improvement after optimization, achieving 91.17% accuracy, while the MLP model only slightly improved from 73.33% to 74.58%. These findings suggest that while synthetic data can support model training, it may also introduce challenges, such as overfitting and difficulties in handling complex medical data. The best models on the small dataset achieved slightly better precision, recall, and f1 scores, indicating that careful management of synthetic data is necessary, especially when dealing with small datasets like the one used in this study, which consisted of only 52 participants [100], [101].

3.5 Image Data

Please refer to Lena Kormacher's paper: *Exploring Mast Cell Diseases with Machine Learning, A Proof of Concept* for more information on the methodology and results of this section.

3.6 Microbiome Analysis

To streamline and standardize the exploratory data analysis (EDA) process, I developed a utility script (utils_MicrobiomeEDA.py). This script contains functions for generating consistent and insightful visualizations and summary statistics, which were used to assess the real microbiome data before data generation and on the augmented data.

3.6.1 Data Augmentation for Microbiome Data

This section details the process of generating synthetic microbiome data for normal individuals and patients with i-MCAS. The Human Microbiome Project data is the basis for these synthetic samples [102]. The objective was to create a dataset that could be used for building and testing ML models.

Methodology:

1. *EDA & Preprocessing:*
   Real microbiome data from the Human Microbiome Project was loaded–The Human Microbiome Project provides a comprehensive dataset of microbial communities from various human body sites. This dataset was chosen due to its high quality and extensive coverage, making it ideal for generating realistic synthetic samples [102].

   The dataset was inspected for basic statistical properties and distributions using functions from the *utils_MicrobiomeEDA.py* script. These functions produced

summary statistics and visualizations such as histograms and box plots to understand the distribution and variability of the data. This step is necessary to identify any anomalies or outliers that might affect the synthetic data generation process. The histograms generated by *utils_MicrobiomeEDA.plot_gene_count_distribution()* revealed most data points concentrated at lower values with a few high-value outliers. The Gene Count histogram is right-skewed, with most counts clustering at lower values, highlighting the frequency of samples with fewer genes (see Figure 37 in Appendix L).

2. *Defining Activators and Suppressors:*
   Based on existing literature, certain bacteria were identified as key players in MC activation or suppression. For instance, bacteria such as Klebsiella pneumoniae and Escherichia coli were considered activators, while Roseburia hominis, Clostridium leptum, and Ruminococcus obeum were considered suppressors [103]. These identifications were critical for accurately simulating potential pathological conditions in i-MCAS patients.

3. *Generating Synthetic Data:*
   The synthetic data generation involved modifying the gene counts of the identified activator and suppressor bacteria. Bootstrapping was employed to generate synthetic samples. Bootstrapping involves sampling with replacement, which creates multiple synthetic datasets by repeatedly resampling the original data. This method was chosen for its strength and ability to create diverse synthetic datasets without introducing substantial biases [93]. For i-MCAS samples, the gene counts of activator bacteria were increased by a factor of 2, while the counts of suppressor bacteria were decreased by a factor of 0.2. This approach ensured that the synthetic i-MCAS samples reflected the intended potential pathological condition.

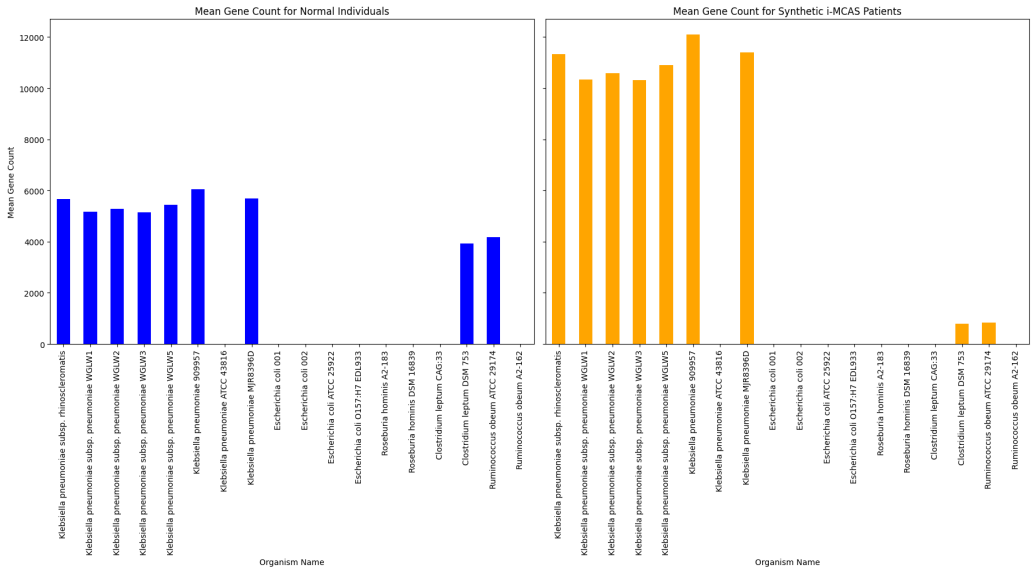
4. *Saving Data:*
   The synthetic datasets for normal and i-MCAS groups were saved as separate CSV files. This format was chosen to make sure each sample set could be kept together and individually identified.

5. *Validation:*
   The synthetic data was validated by comparing the mean gene counts of selected bacteria between the normal and i-MCAS groups. This comparison was done using statistical tests such as t-tests to confirm that differences were achieved and the data reflected the intended augmentation (see Figure 35 in Appendix K). A bar plot was used to present these differences.

Results:
The synthetic microbiome data for normal and i-MCAS samples was successfully generated using the Human Microbiome Project data as a basis. Validation showed distinct differences in the mean gene counts of activator and suppressor bacteria between normal and i-MCAS samples (see Figure 10). The generated synthetic data accurately

reflected the intended potential pathological condition, confirming the effectiveness of the data augmentation process.

Figure 10: *Mean Gene Counts of Activator and Suppressor Bacteria*



Source: Lea Brody-Heine, *Microbiome Data/Microbiome Data Aug.ipynb*

The visualizations provided clear evidence of these differences, with the bar plots showing higher gene counts for activator bacteria and lower counts for suppressor bacteria in i-MCAS samples compared to normal samples.

3.6.2 Microbiome Analysis Model

This section focuses on analyzing synthetic microbiome data to identify patterns and differences between a control group and a group with i-MCAS. The goal is to understand the microbiome composition's possible role in i-MCAS and identify specific bacteria that differ greatly between the two groups. The analysis involves several key steps feature extraction, normalization, dimensionality reduction, clustering, and detailed cluster analysis.

Methodology:

1. *EDA & Preprocessing:*
   Synthetic microbiome data was loaded from multiple CSV files and combined into a single DataFrame. Each sample was assigned a unique Sample ID for identification reasons. This approach guaranteed all samples were consolidated correctly for easier analysis and processing. The data was pivoted to have organisms as columns and samples as rows, transforming the dataset into a format suitable for ML models. Missing values were filled with zeros to ensure completeness. Standard scaling was applied to normalize the data, ensuring all features contributed equally to the model. This step prevented any feature from dominating due to scale differences.

   The EDA from *utils_MicrobiomeEDA.py* showed a predominance of bacterial organisms, significant variation in microbial counts across body sites, particularly

in the gastrointestinal tract and oral sites, identified the top ten most frequent organisms, among other insights. The output from the augmented data matched the real data during exploration indicating a successful augmentation capturing the patterns of the real data (see Figures in Appendix L).

2. *Dimensionality Reduction With Principal Component Analysis (PCA):*
   PCA was used to reduce the dimensionality of the data to two principal components, simplifying the dataset while retaining most of the variance. This made it easier to visualize and identify patterns or groupings in the data. Principal components were inspected and analyzed to show the top contributing features of each principal component. Subsequently, this visualized to understand which variables (bacteria) drove the variance in the data (see Figure 43 in Appendix M).

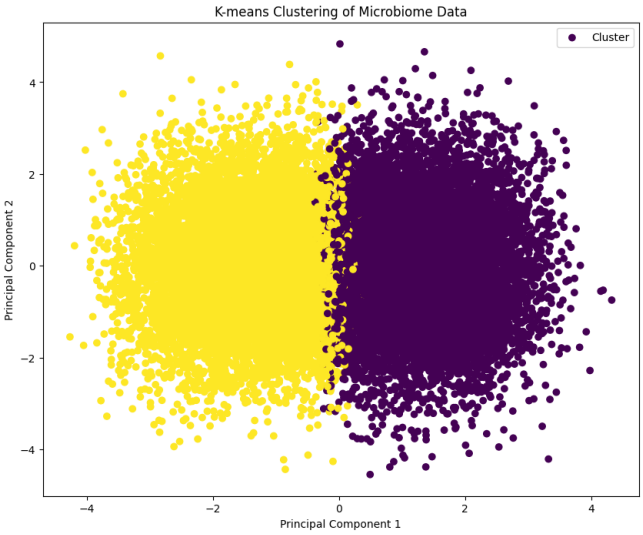
3. *Unsupervised Learning Model:*
   K-Means and GMM clustering were applied to the normalized data to identify distinct groups within the samples. The clusters were visualized in the PCA-reduced space to understand the natural groupings in the data. K-Means clustering is a simple yet effective method for identifying distinct groups within the data. GMM Clustering was used to identify meaningful differences between i-MCAS and normal samples using a probabilistic approach. For both methods, cluster labels were added to the data, mean values for each feature within each cluster were calculated, and the top features with the highest differences between clusters were determined. This analysis revealed which bacteria were notably more prevalent in either the i-MCAS or the normal group, providing insights into potential microbial influences on the condition.

Results:
The synthetic data was used to successfully build an ML pipeline for microbiome analysis. Validation confirmed differences between the normal and i-MCAS groups, aligning with expected pathological conditions.

The scatter plots in PCA-reduced space for both K-Means and GMM clustering showed clear groupings, indicating that these clustering methods effectively captured the natural groupings within the data, reflecting the underlying differences in microbiome composition between the two groups (see Figure 11 and 44, 45 in Appendix M). The use of both K-Means and GMM clustering strengthens the pipeline by cross-validating the clustering results. Because both models produced similar clusters, there is further confidence in the findings.

Both clustering methods confirmed two distinct clusters corresponding to the normal and i-MCAS groups. The cluster analysis from both K-Means and GMM models revealed that certain bacteria, such as strains of Klebsiella pneumoniae and Escherichia coli, were more prevalent in i-MCAS samples, while Ruminococcus obeum and Clostridium leptum were more common in normal samples (see Figure 12 and 45 in Appendix M). These results show the potential for discovering bacterial biomarkers associated with i-MCAS. The identified bacteria matched the expected outcomes when cross-checked with the original augmentation activators and suppressors.

Figure 11: *K-Means Clustering*

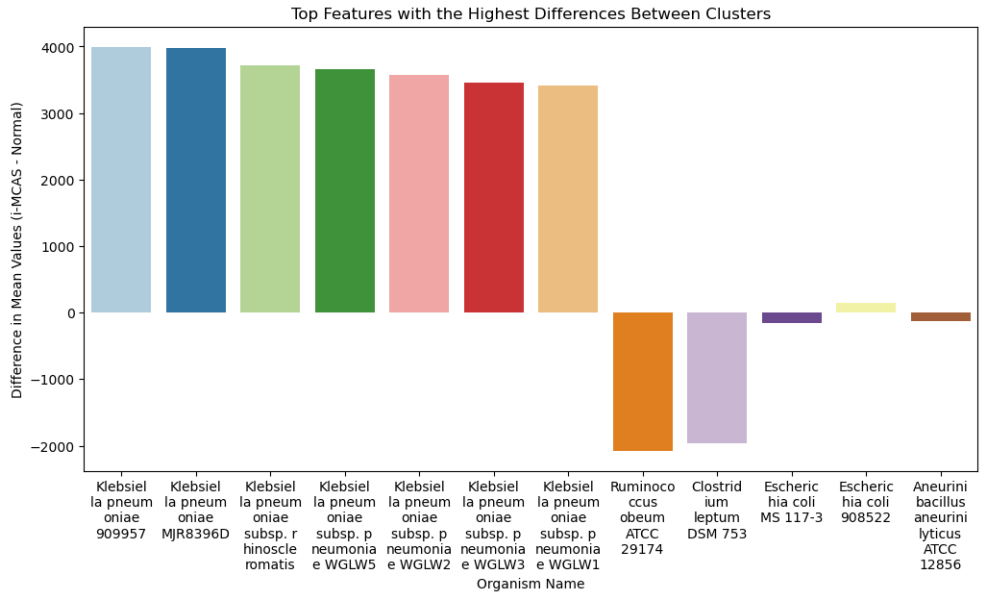


Source: Lea Brody-Heine *Microbiome Models.ipynb*

This alignment suggests the model performed accurately, providing a reliable tool for further studies.

The convergence of results from two different clustering algorithms suggests that the identified bacterial differences are not artifacts of a single model's assumptions but are more likely to reflect true biological differences. This redundancy in modeling enhances the reliability of the analysis, offering a stronger foundation for discovering potential bacterial biomarkers associated with i-MCAS. The identified bacteria matched the expected outcomes when cross-checked with the original augmentation activators and suppressors, further validating the effectiveness of the models and the overall pipeline.

Figure 12: *Cluster Analysis*



Source: Lea Brody-Heine, *Microbiome Models.ipynb*

3.7 Model Application

This section outlines the model application processes in *'script.py'*. We focus on preprocessing new data, applying saved models for prediction, and finally, integrating image classification and object detection.

1. Preprocessing of New Data:
   The preprocessing of new data is necessary to confirm the input data is in a

suitable format for prediction by the trained models. The preprocessing script includes several key steps:

(a) The new data is loaded from a CSV file into a Pandas DataFrame. Any empty strings in the dataset are replaced with NaN values to handle missing data appropriately.

(b) Numerical features are identified by selecting columns with numeric data types. These features are then converted to numeric types, with any errors in conversion resulting in NaN values.

(c) If the ”Diagnosis” column exists in the new data, it is dropped to prevent any leakage of target information during prediction.

(d) The custom preprocessing pipeline, defined in a separate utility script *utils_ApplyPipeline*, is applied to the data to maintain standardized preprocessing and consistency of tabular data throughout the project before model prediction or training.

2. Tabular Model Application and Prediction:
Script.py includes functions to apply saved models for generating predictions. The key steps are outlined as follows:

(a) The new data's initial shape and column count are printed for verification. The saved XGBoost and MLP models are loaded using the joblib library.

(b) Features are extracted from the new data using the preprocess_new_data function, which applies the preprocessing pipeline and handles any necessary data cleaning and transformation.

(c) Predictions are made using both the XGBoost and MLP models. The predicted values are added to the DataFrame as new columns ('XGBoost_Prediction' and 'MLP_Prediction').

(d) The predictions are printed to verify the output and confirm the models function as expected.

3. Image Classification & Object Detection: Please see Lena Korfmacher's paper: *Exploring Mast Cell Diseases with Machine Learning, A Proof of Concept* for an explanation of application methodology.

Results:
The script combines the main components of our project and provides users with an easy way to make predictions by applying the trained and tested models to new data. The script requires specifying a path to a csv file that includes the new tabular data, a path to a new image, and, if necessary, selecting a suitable model.

After preprocessing, the script used the XGBoost and MLP models to make predictions on this data. The output showed that the XGBoost model predicted various classes for the data points, with predictions of 0, 2, 0, 1, and 0 for the five entries. In contrast, the MLP model consistently predicted class 2 for all entries, indicating a potential issue with the MLP model's ability to differentiate between the classes in this dataset. The discrepancy between the models' predictions suggests that the XGBoost

model handled the data more effectively, whereas the MLP model may require further tuning or adjustments to better capture the data's nuances.

Currently, the tabular data does not correspond to the biopsy image. When using the application in the future, care should be taken to connect the tabular data and biopsy images from the same patient to enable consistency in patient pathology. Further, the image classifier model was too large to save and push into GitHub. Therefore, the image classification code in the script serves as a placeholder. Future implementations will require different technologies with more available memory to handle the full model and seamlessly integrate the tabular data. Please see Lena Kormacher's paper: *Exploring Mast Cell Diseases with Machine Learning, A Proof of Concept* for more information regarding image classification and object detection.

The microbiome model was not included here because it is not a classification model nor a supervised learning model. Instead, it focuses on identifying patterns and making discoveries for research purposes, which do not align well with the current application file's structure.

4 Discussion & Critical Appraisal

Our dissertation presents a proof-of-concept approach to exploring and diagnosing MCDs using various ML and DL algorithms. The work does not claim to provide scientific evidence but instead offers a framework that could be replicated and expanded upon with more comprehensive data. A significant limitation of this research is the insufficient data underpinning, as ML and DL models require extensive data to learn and improve.

Despite these limitations, our research demonstrates that ML models can classify two MC disorders, HαT and MCAS. The models can also distinguish the two diagnoses from the control group subjects with other gastrointestinal conditions.

The innovative approach in this dissertation, along with our integration of diverse data sources and ML/DL techniques, provides results that are not reliant on a single AI outcome or the judgment of one human expert. While the primary objective was to develop tools that could support researchers studying MCDs and medical diagnostics, the artifact underscores the need for further work to refine these methods and validate them with more extensive, high-quality data.

Our work highlights the potential of ML and DL in the field of MCD diagnosis but also points to the need for replication and further development to achieve the level of reliability required for clinical and research applications.

4.1 Tabular Data

4.1.1 Strengths, Limitations, and Challenges

Using both the original and augmented datasets provided distinct strengths and challenges. Some of the models trained on the original dataset, although small with only 52 participants, had relatively high accuracy demonstrating that even limited data can yield meaningful insights with the correct algorithm application. However, the small

sample size also presented challenges, particularly for more complex models like MLP, which generally require larger datasets to capture underlying patterns effectively.

The tabular data required extensive preprocessing and transformation to be suitable for model training. Handling longitudinal data and multiple procedures per patient added complexity, necessitating careful data management to provide the models with accurate and relevant information. Despite these efforts, the small size of the dataset limited more complex models' ability to generalize, highlighting the need for more data to improve these models' performance.

The augmentation process aimed to address the limitations of the small dataset by increasing its size and introducing synthetic variability. This provided more data for training and helped the models generalize better, as seen in the slight improvement in MLP accuracy. Exploring augmented data can reveal important patterns not present in the original dataset. Nevertheless, it should be noted that the approach used for augmenting tabular data was not state-of-the-art, as more advanced methods like those discussed by Murtaza et al. could potentially have produced better results [104]. The use of GANs for data augmentation was unsuccessful, as the generated synthetic data failed to mimic the original dataset accurately, further contributing to the challenges in achieving effective model training.

There were considerable limitations with the augmented dataset. Introducing synthetic variability also introduced noise, leading to reduced model performance compared to the original dataset. This was particularly evident in the XGBoost and MLP models, where the augmented data led to overfitting and decreased generalizability [101]. Additionally, techniques like SMOTE, intended to address class imbalances, were not effective and resulted in biased predictions, especially for the normal class [22], [99].

Training models on both datasets presented challenges. The complexity of medical data means that subtle and important patterns can be lost or misrepresented, particularly when augmentation introduces noise. The small sample size of the original dataset limited the effectiveness of the models, while the augmented dataset's synthetic variability did not fully address these limitations. Proper handling of longitudinal data and multiple procedures per patient remained a challenge, as did the non-ideal augmentation approach. Predictions on test image data were not made on truly unseen data, as some synthetic versions of the test images were already present in the training and validation sets due to a lack of original biopsy images. This further complicated the assessment of model performance.

While the augmented dataset provided some benefits, it also introduced new challenges that highlight the need for more advanced augmentation techniques and mindful data management in future research.

4.1.2 Broader Implications

The scripting applied in this study highlights the potential for integrating various ML models into a cohesive framework that can be used to predict outcomes based on tabular and image data. The ability to load new data, preprocess it, and generate predictions using saved models like XGBoost and MLP is particularly useful for clinical and research applications. However, the discrepancies between the performance of

models trained on original and augmented data emphasize the need for careful consideration of data quality and model training techniques in future applications.

The marginal gains in MLP performance when using augmented data suggest that not all models benefit equally from data augmentation, particularly when synthetic data introduces noise or fails to capture the complexity of real-world patterns. This finding underscores the importance of selecting appropriate augmentation techniques and ensuring that synthetic data closely mimics the characteristics of the original dataset.

Furthermore, the broader implications of this study extend to general ML models in medical applications. The success of the XGBoost model in maintaining high accuracy across both original and augmented datasets indicates that tree-based models may be more resilient to variations in data quality. This suggests that XGBoost and similar algorithms could be prioritized in scenarios where data augmentation is necessary, but model robustness remains a concern.

The integration of these ML models into a streamlined workflow, as demonstrated in this study, provides a foundation for developing more advanced diagnostic and research tools. By refining augmentation techniques and optimizing model performance, future research can build upon this framework to create reliable, scalable solutions for analyzing complex medical data. Ultimately, this approach has the potential to enhance the accuracy and efficiency of medical research, leading to better patient outcomes, more informed clinical decision-making, and better treatments.

4.1.3 Objectives Met

The tabular data section met Objectives 1, 2, and 3 across both the original and augmented datasets (please see Lena Kormacher's paper: *Exploring Mast Cell Diseases with Machine Learning, A Proof of Concept* for information on the success of image data for Objectives 2 & 3).

Objective 1: Develop a ML Algorithm for Pathological Data Analysis:
For the original data, the XGBoost model demonstrated strong performance, showing that traditional ML algorithms can effectively analyze pathological data. Yet, the performance slightly dropped when training on the augmented dataset, even after hyperparameter optimization.

Objective 2: Explore Generative AI Models for Synthetic Data Generation:
We successfully implemented Objective 2 and were able to compare augmentation techniques and found the traditional augmentation to increase the dataset size proved more effective at capturing the patterns from the real dataset. The GAN-based approach failed to produce a useful dataset as it had difficulty creating high-quality synthetic data that closely mimicked the real set. The traditional augmentation methods, despite their limitations, were more successful in supporting the training process compared to the GAN approach. As a result, Objective 2 was met but it also emphasizes the challenges posed by synthetic data generation stressing the need for further refinement and alternative strategies in future work.

Objective 3: Develop a Proof-of-Concept AI Tool for Clinical Application and Re-

search:
The development of ML models that integrated both original and augmented data met the criteria for Objective 3. These models demonstrated the potential for creating a scalable AI tool suitable for clinical and research purposes. The project succeeded in highlighting both the strengths and limitations of these approaches, especially in the context of MCDs.

4.2 Image Data

Please see *Lena Korfmacher's paper: Exploring Mast Cell Diseases with Machine Learning, A Proof of Concept* for the discussion on the image data, methodology, and results.

4.3 Microbiome Analysis

4.3.1 Strengths, Limitations, and Challenges

One of the main strengths of this section is using synthetic data to develop and test methods for microbiome research. This approach created a flexible framework that can be applied to real patient data. The methods used, including feature extraction, PCA, K-means clustering, and Gaussian Mixture Model (GMM) clustering, provide a comprehensive model for microbiome analysis [105]. The inclusion of GMM alongside K-means clustering offers an additional layer of validation because both models independently identified similar clusters within the data. This congruence between two distinct clustering algorithms reinforces the strength of the identified groupings. It indicates that the underlying patterns in the microbiome data are well-captured by the chosen models. In future work, employing additional clustering models would be advisable. I could not explore these options due to the limitations of GPU support and the time constraints associated with running such models. Still, future use of advanced dimensionality reduction techniques like t-SNE could make this more feasible [22]. Lastly, my pipeline ensures reproducibility and consistency, making it a valuable tool for future research into the microbiome's role in health and disease.

The microbiome analysis section has several limitations. The augmentation process relies on a single study to identify MC activators and suppressors, which may not capture the full complexity of microbial and immune system interactions. Using gene count as the measure of abundance data may not be the most accurate representation; future research should incorporate more precise abundance measures. Gene counts can be misleading due to variations in 16S rRNA gene copy numbers among different taxa, which can overestimate or underestimate actual organismal abundance (see Appendix C for more information). Normalization strategies, such as Quantitative Microbiome Profiling (QMP) or using reference taxa, provide more accurate abundance measures by adjusting for these variations [106]–[108]. Likewise, the current model cannot determine whether the observed microbiome disparities are a cause of i-MCAS or merely another symptom of the condition.

Additionally, the models recognize only binary outcomes due to the binary data, whereas real data will likely present more complex clustering beyond merely two groups. This simplification does not fully capture the potential diversity in actual patient micro-

biome compositions. While synthetic data provides a controlled environment for model training, the nuances of real-world data will likely lead to different outcomes and insights. Efforts must improve the models' robustness to guarantee it will accommodate a wider range of microbiome diversity in future analyses.

Several challenges were encountered during the microbiome analysis. Augmenting data was especially difficult because each sample represented an entire CSV file of microbial data. Integrating this large amount of information into a format suitable for model training was complicated. For example, a single person's sample could contain data on 2,000-3,000 different microbes with each microbe represented as a single row. The integrity of each sample's data needed to be maintained but also structured correctly for analysis. Developing a methodology to handle this complexity was a critical part of the section's success but also one of the most challenging.

4.3.2 Broader Implications

The findings from the microbiome analysis have several broader implications for both scientific research and clinical practice. These implications are beyond the immediate results of this dissertation and suggest avenues for future exploration and potential real-world applications

This dissertation demonstrates the potential of using synthetic microbiome data to train ML models. By providing a framework that can be applied to real patient data, researchers will be equipped with the tools necessary to analyze microbiome compositions in i-MCAS and normal individuals. This approach identifies microbial patterns and biomarkers that could be critical in understanding the underlying mechanisms of i-MCAS.

The significance of this dissertation is in applying these analytical methods to actual patient data. Researchers can plug their own microbiome datasets into the established pipeline, leveraging the existing model and analysis techniques without needing to rebuild pipelines or outsource ML analysis. This flexibility enables future studies to efficiently gain insights from their specific datasets.

The methodologies employed, including feature extraction, PCA, and K-Means clustering, provide a standardized analytical method researchers can use to maintain consistency and comparability across different studies. Therefore, enhancing the reproducibility and scalability of i-MCAS microbiome research. This could contribute to a more comprehensive understanding of how the microbiome influences health and disease [109], [110].

If validated with real patient data, non-invasive diagnostic tests based on microbiome profiles could be created, improving patient outcomes through timely and accurate diagnosis. Additionally, personalized medicine approaches could benefit from this research, as understanding the unique microbial profiles of i-MCAS patients could lead to targeted treatments that modulate the microbiome. Thus, resulting in more effective therapies and improved quality of life for patients [111], [112].

The broader implications of this dissertation emphasize the importance of continued research into the microbiome's role in health and disease. By advancing scientific knowledge and providing practical tools for researchers, this dissertation has the potential to impact the understanding and treatment of i-MCAS and other related condi-

tions.

4.3.3 Objectives Met

Objective 4: Build a tool for the Investigation of Microbiome Discrepancies in i-MCAS Patients:
This section of the dissertation successfully met objective 4. The implementation of data augmentation techniques, including bootstrapping, generated comprehensive microbiome datasets for both normal individuals and i-MCAS patients. I manipulated specific bacteria classes to emulate potential discrepancies in these groups and ensure the synthetic data accurately represented the expected microbial imbalances. An unsupervised learning model was used to identify and characterize significant microbiome discrepancies between i-MCAS patients and healthy individuals. PCA and K-Means clustering created visualizations to identify distinct microbial patterns and differences. These analyses confirmed the presence of potential differences in microbial compositions between the two groups, aligning with the expected pathological conditions and validating the approach used in this dissertation.

4.4 Future Work & Recommended Use

Based on the findings, we recommend focusing future work on applying the models to larger, high-quality datasets. In clinical or research settings with ample data, the need for data augmentation, as used in this study to supplement the lack of data, would be minimized. Acquiring and utilizing comprehensive, clean datasets to enhance model accuracy and reliability without relying on synthetic data should be a priority.

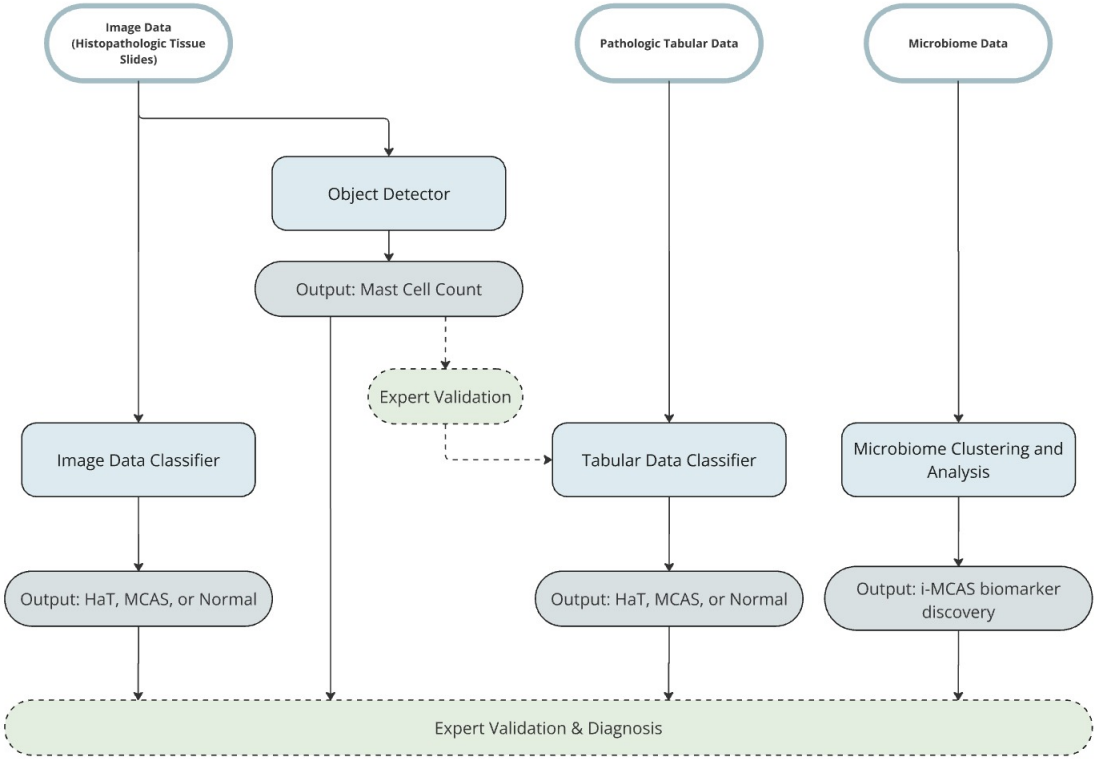
The models showed promise, especially after hyperparameter optimization, but continued validation with expert clinicians will be essential to confirm the models' effectiveness. The framework developed in this dissertation offers a solid starting point, noting that collaboration with clinical professionals and iterative testing will be necessary for fine-tuning these models to meet diagnostic needs and fully develop their potential.

In the idealized application flow presented in Figure 13, we outline a process for integrating our ML models to support both diagnosis and research in MCDs. The process starts by using image data from histopathologic tissue slides that are processed through an object detection model to count MCs. This count should be validated by an expert before being fed into a tabular data classifier that categorizes the samples into the three diagnostic classes. The count could also be used on its own as further information about a patient's health.

Pathologic tabular data would be processed through a tabular data classifier, also yielding an output of HαT, MCAS, or normal. Expert validation will play a crucial role at this stage as well to confirm that the classifier's output aligns with clinical expectations.

Separately, microbiome data could be analyzed using K-Means clustering to discover novel biomarkers specific to i-MCAS. This analysis would not feed directly into the diagnostic process but would be used for research purposes to enhance the understanding of i-MCAS and potentially uncover new therapeutic targets.

Figure 13: *Suggested Integration of Modules (idealized)*



Source: Lea Brody-Heine & Lena Korfmacher — Miro

4.4.1 Utilize The Code Pipelines For Real Data

To run the project and analyze your own data, follow these steps:

Upload Your Data:

- Prepare your data in a format consistent with the project's existing datasets (e.g., CSV for tabular data with similar features).
- Upload the data to the appropriate directory, such as 'Tabular Data/' or 'Image Data/'.

Specify Data Paths:
Open the relevant notebook and update the data loading commands to point to your files. For example:

```
data = pd.read_csv('Tabular Data/your_data_file.csv')
```

Choose a Pipeline:

- Select an appropriate pipeline from the provided notebooks for data preprocessing, augmentation, and model training.
- Customize the pipeline if necessary by modifying preprocessing steps or model parameters.

Run the Pipeline:
Execute the pipeline to clean, preprocess, and augment your data as needed. This includes tasks like handling missing values, scaling, and encoding.

Train and Evaluate Models:

- Split your data into training and testing sets, then train the model using the chosen pipeline.
- Evaluate the model's performance and save the trained model for future use.

Use the Trained Model for Predictions:

- Save your trained model (e.g., as a '.pkl' file).
- Use the 'script.py' file to load the saved model and make predictions on new data by specifying the model and data paths in the script.

5 Ethical Considerations

This dissertation was approved by the Ethics Committee of the University of St Andrews. Please see appendix D for the approval document.

We have carefully considered the broader ethical implications of applying AI and ML technologies in the medical field. The sensitive nature of our obtained medical data necessitates stringent data privacy measures, particularly in the context of machine learning models that require large datasets for training. While the data was anonymized before our use, we maintained its secure storage as a critical measure for preserving confidentiality and protecting patient rights.

Potential bias in AI models, especially when dealing with medical diagnostics, is a important ethical concern. Biases in training data can lead to unequal treatment outcomes, disproportionately affecting underrepresented groups [113]. In future work, these concerns must be addressed by implementing strategies to identify and mitigate biases during model development.

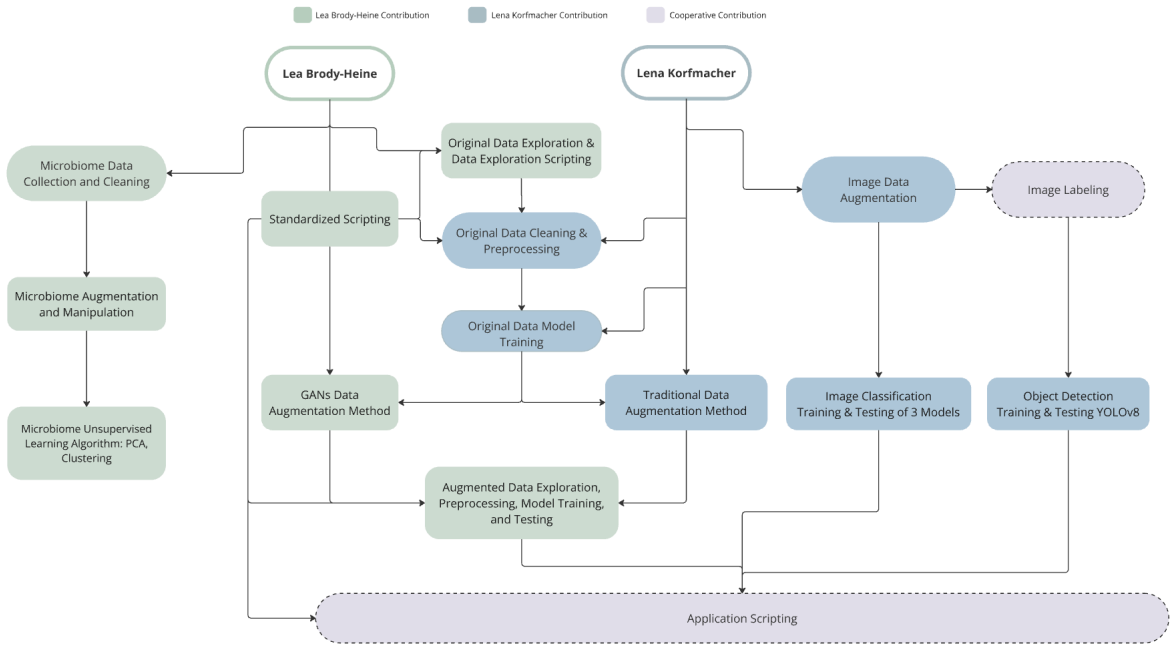
The transparency and interpretability of AI models are critical in clinical settings. Clinicians and researchers must understand and trust the decisions made by these systems, especially when they directly impact patient care. This research emphasizes the further need for xAI methods to guarantee that the outputs of the models are interpretable and actionable by medical professionals.

6 Summary

This dissertation explored the use of machine learning techniques to tackle the diagnostic and research challenges associated with mast cell diseases, specifically idiopathic mast cell activation syndrome (i-MCAS) and hereditary alpha-tryptasemia (HαT). By integrating diverse data types–pathological data, microbiome samples, and biopsy images–we developed and evaluated models that can identify patterns that traditional methods might overlook. To overcome the limitation of small datasets, which

is a common issue in medical research, we compared traditional data augmentation to generative models and found more success with traditional augmentation.

Figure 14: *Contribution Flow Chart*



Source: Lea Brody-Heine & Lena Korfmacher — Miro

Our approach is relatively novel, particularly in the context of combining machine learning with microbiome analysis and histopathological image data to study mast cell diseases. The current literature often focuses on these elements in isolation, but the integration seen in this work represents a more holistic approach that could lay the groundwork for more accurate and comprehensive diagnostic tools. While the results were promising, there are areas for improvement, especially regarding model fine-tuning and tackling the ethical implications of bias in AI models. Our dissertation contributes to the growing body of research that seeks to leverage AI for complex medical diagnostics and research, demonstrating both its potential and the need for continued exploration and refinement.

7 Personal Contribution

- Tabular Data Exploration Scripting and development
- Model Training on the Augmented Data
- Entire Microbiome Section
- GAN Data Augmentation Attempt
- Responsible for scripting, refactoring, and standardization of processes and methodologies through the code base
- Built the model application script: script.py

References

[1] P. M. Leru, "Evaluation and classification of mast cell disorders: A difficult to manage pathology in clinical practice," *Cureus*, vol. 14, no. 2, Feb. 2022. DOI: 10.7759/cureus.22177.

[2] N. Malik, U. Junaid, and A. Liaqat, "The gut microbiome and its influence on immune function: A comprehensive review," *International Journal of Advanced Engineering Technologies and Innovations*, vol. 3, no. 1, 2024. [Online]. Available: https://ijaeti.com/index.php/Journal/article/view/336.

[3] M. J. Hamilton, M. Zhao, M. P. Giannetti, *et al.*, "Distinct small intestine mast cell histologic changes in patients with hereditary alpha-tryptasemia and mast cell activation syndrome," *The American journal of surgical pathology*, vol. 45, no. 7, pp. 997–1004, Jul. 1, 2021, ISSN: 0147-5185. DOI: 10.1097/PAS.0000000000001676. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8192345/ (visited on 05/23/2024).

[4] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959. DOI: 10.1147/rd.33.0210.

[5] T. M. Mitchell, *Machine Learning* (McGraw-Hill series in computer science). New York: McGraw-Hill, 1997, 414 pp., ISBN: 978-0-07-042807-2.

[6] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nature Methods*, vol. 15, no. 4, pp. 233–234, 2018. DOI: 10.1038/nmeth.4642.

[7] C. University. "Artificial intelligence (ai) vs. machine learning." CU-CAI. Retrieved June 27, 2024, from https://ai.engineering.columbia.edu/ai-vs-machine-learning/. (n.d.).

[8] M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, Article 5, 2023. DOI: 10.3390/computers12050091.

[9] IBM. "Deep blue — ibm." Retrieved June 25, 2024, from https://www.ibm.com/history/deep-blue. (n.d.-a).

[10] IBM. "Ibm watson." Retrieved June 25, 2024, from https://www.ibm.com/watson. (n.d.-b).

[11] G. DeepMind. "Alphago." Retrieved May 14, 2024, from https://deepmind.google/technologies/alphago/. (2024).

[12] K. Sharifani and M. Amini, "Machine learning and deep learning: A review of methods and applications," vol. 10, no. 07, 2023.

[13] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021. DOI: 10.1007/s42979-021-00592-x.

[14] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 2022. DOI: 10.1038/s41591-021-01614-0.

[15] A. R. Kunduru. “Machine learning in drug discovery: A comprehensive analysis of applications, challenges, and future directions.” Retrieved August 2023, from https://dspace.umsida.ac.id/handle/123456789/16754. (2023).

[16] C. J. Haug and J. M. Drazen, “Artificial intelligence and machine learning in clinical medicine, 2023,” *New England Journal of Medicine*, vol. 388, no. 13, pp. 1201–1208, 2023. DOI: 10.1056/NEJMra2302038.

[17] Z. Yao, Y. Lum, A. Johnston, *et al.*, “Machine learning for a sustainable energy future,” *Nature Reviews Materials*, vol. 8, no. 3, pp. 202–215, 2023. DOI: 10.1038/s41578-022-00490-5.

[18] S. J. Qin and L. H. Chiang, “Advances and opportunities in machine learning for process data analytics,” *Computers & Chemical Engineering*, vol. 126, pp. 465–473, 2019. DOI: 10.1016/j.compchemeng.2019.04.003.

[19] IBM. “What is artificial intelligence (ai)?” Retrieved August 25, 2023, from https://www.ibm.com/topics/artificial-intelligence. (2023).

[20] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, “A proposal for the dartmouth summer research project on artificial intelligence,” Aug. 31, 1955.

[21] A. M. Turing, “Computing Machinery and Intelligence,” *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. [Online]. Available: https://doi.org/10.1093/mind/LIX.236.433 (visited on 06/21/2024).

[22] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media, Incorporated, 2022, http://ebookcentral.proquest.com/lib/st-andrews/detail.action?docID=30168989.

[23] L. Alzubaidi, J. Zhang, A. J. Humaidi, *et al.*, “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, ISSN: 2196-1115. DOI: 10.1186/s40537-021-00444-8. [Online]. Available: https://doi.org/10.1186/s40537-021-00444-8 (visited on 08/03/2024).

[24] IBM, *What is Computer Vision?* en-us, Jul. 2021. [Online]. Available: https://www.ibm.com/topics/computer-vision (visited on 07/27/2024).

[25] S. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, Conference Name: IEEE Transactions on Information Theory, ISSN: 1557-9654. DOI: 10.1109/TIT.1982.1056489. [Online]. Available: https://ieeexplore.ieee.org/document/1056489/?arnumber=1056489 (visited on 08/03/2024).

[26] K. P. Sinaga and M.-S. Yang, “Unsupervised K-Means Clustering Algorithm,” *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988796. [Online]. Available: https://ieeexplore.ieee.org/document/9072123/?arnumber=9072123 (visited on 08/03/2024).

[27] D. O. Hebb, *The Organization of Behavior*. JOHN WILEY & SONS, Inc., 1949, https://pure.mpg.de/rest/items/item_2346268_3/component/file_2346267/content.

[28] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” en, *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, ISSN: 1522-9602. DOI: 10.1007/BF02478259. [Online]. Available: https://doi.org/10.1007/BF02478259 (visited on 07/04/2024).

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” in *Neurocomputing, Volume 1*, J. A. Anderson and E. Rosenfeld, Eds., Cambridge, MA: The MIT Press, 1986, pp. 318–362, ISBN: 978-0-262-26713-7. DOI: 10.7551/mitpress/4943.003.0128. [Online]. Available: https://direct.mit.edu/books/book/5431/chapter/4468772/1986-D-E-Rumelhart-G-E-Hinton-and-R-J-Williams (visited on 07/04/2024).

[30] K. Suzuki, *Artificial Neural Networks: Architectures and Applications*. BoD – Books on Demand, Jan. 2013, Google-Books-ID: JUufDwAAQBAJ, ISBN: 978-953-51-0935-8.

[31] Z. Meng, Y. Hu, and C. Ancey, “Using a Data Driven Approach to Predict Waves Generated by Gravity Driven Mass Flows,” *Water*, vol. 12, Feb. 2020. DOI: 10.3390/w12020600.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3422622. [Online]. Available: https://dl.acm.org/doi/10.1145/3422622 (visited on 07/28/2024).

[33] A. Loth, M. Kappes, and M.-O. Pahl, *Blessing or curse? A survey on the Impact of Generative AI on Fake News*, arXiv:2404.03021 [cs], Apr. 2024. [Online]. Available: http://arxiv.org/abs/2404.03021 (visited on 07/27/2024).

[34] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1249, 2018, ISSN: 1942-4795. DOI: 10.1002/widm.1249. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249 (visited on 08/05/2024).

[35] F. Zhuang, Z. Qi, K. Duan, *et al.*, “A Comprehensive Survey on Transfer Learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, Conference Name: Proceedings of the IEEE, ISSN: 1558-2256. DOI: 10.1109/JPROC.2020.3004555. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9134370/references#references (visited on 08/04/2024).

[36] A. Halevy, P. Norvig, and F. Pereira, “The Unreasonable Effectiveness of Data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009, Conference Name: IEEE Intelligent Systems, ISSN: 1941-1294. DOI: 10.1109/MIS.2009.36. [Online]. Available: https://ieeexplore.ieee.org/document/4804817/?arnumber=4804817 (visited on 08/02/2024).

[37] M. Banko and E. Brill, “Scaling to very very large corpora for natural language disambiguation,” in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL ’01*, Toulouse, France: Association for Computational Linguistics, 2001, pp. 26–33. DOI: 10.3115/1073012.1073017. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1073012.1073017 (visited on 08/02/2024).

[38] G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd, "Training set size requirements for the classification of a specific class," *Remote Sensing of Environment*, vol. 104, no. 1, pp. 1–14, Sep. 2006, ISSN: 0034-4257. DOI: 10.1016/j.rse.2006.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425706001234 (visited on 07/28/2024).

[39] C. A. Ramezan, T. A. Warner, A. E. Maxwell, and B. S. Price, "Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data," *Remote Sensing*, vol. 13, no. 3, p. 368, Jan. 2021, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-4292. DOI: 10.3390/rs13030368. [Online]. Available: https://www.mdpi.com/2072-4292/13/3/368 (visited on 07/28/2024).

[40] GeeksforGeeks, *How much data is sufficient to train a machine learning model?* en-US, Section: AI-ML-DS, Feb. 2024. [Online]. Available: https://www.geeksforgeeks.org/how-much-data-are-sufficient-to-train-my-machine-learning-model/ (visited on 07/28/2024).

[41] V. Berger and Y. Zhou, "Kolmogorov–smirnov test: Overview," in *Wiley StatsRef: Statistics Reference Online*, N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. Teugels, Eds., Wiley, 2014. DOI: 10.1002/9781118445112.stat06558. [Online]. Available: https://doi.org/10.1002/9781118445112.stat06558.

[42] M. Talo, "Automated classification of histopathology images using transfer learning," *Artificial Intelligence in Medicine*, vol. 101, p. 101 743, Nov. 2019, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2019.101743. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365719307110 (visited on 07/05/2024).

[43] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," en, *BMC Medical Imaging*, vol. 22, no. 1, p. 69, Apr. 2022, ISSN: 1471-2342. DOI: 10.1186/s12880-022-00793-7. [Online]. Available: https://doi.org/10.1186/s12880-022-00793-7 (visited on 08/04/2024).

[44] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable ai methods—a brief overview," in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., Springer International Publishing, 2022, pp. 13–38. DOI: 10.1007/978-3-031-04083-2_2.

[45] GDPR.eu, *What is GDPR, the EU's new data protection law?* Section: GDPR Overview, Nov. 2018. [Online]. Available: https://gdpr.eu/what-is-gdpr/ (visited on 07/27/2024).

[46] E. Commission, *AI Act — Shaping Europe's digital future*, en, Jun. 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (visited on 07/16/2024).

[47] H. Smith, "Clinical AI: Opacity, accountability, responsibility and liability," *AI & SOCIETY*, vol. 36, no. 2, pp. 535–545, Jun. 2021, ISSN: 1435-5655. DOI: 10.1007/s00146-020-01019-6. [Online]. Available: https://doi.org/10.1007/s00146-020-01019-6 (visited on 08/03/2024).

[48] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, no. 1, 2021. DOI: 10.3390/e23010018.

[49] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. DOI: 10.1016/j.artint.2018.07.007.

[50] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” *AI Magazine*, vol. 40, no. 2, 2019. DOI: 10.1609/aimag.v40i2.2850.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, ISSN: 2380-7504, Oct. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74. [Online]. Available: https://ieeexplore.ieee.org/document/8237336/?arnumber=8237336 (visited on 08/04/2024).

[52] A. Hedström, L. Weber, D. Bareeva, *et al.*, “Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond,” *arXiv*, 2023. DOI: 10.48550/arXiv.2202.06861.

[53] L. Arras, A. Osman, and W. Samek, “Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations,” *Information Fusion*, vol. 81, pp. 14–40, 2022. DOI: 10.1016/j.inffus.2021.11.008.

[54] M. Fong and J. S. Crane, “Histology, mast cells,” 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK499904/ (visited on 06/20/2024).

[55] “Overview & diagnosis,” TMS - The Mast Cell Disease Society, Inc. (), [Online]. Available: https://tmsforacure.org/overview/ (visited on 06/20/2024).

[56] *The mast cell disease primer*, 2021. [Online]. Available: https://tmsforacure.org/wp-content/uploads/Mast_Cell_Disease_Primer_Slides_TMS_09.20.2021_Final.pdf.

[57] R. D. Shamburek and M. L. Schubert, “Control of gastric acid secretion. histamine h2-receptor antagonists and h+k(+)-atpase inhibitors,” *Gastroenterology clinics of North America*, vol. 21, no. 3, pp. 527–550, 1992.

[58] D. Khokhar and C. Akin, “Mast cell activation,” *Medical Clinics of North America*, vol. 104, no. 1, pp. 177–187, Jan. 2020. DOI: 10.1016/j.mcna.2019.09.002.

[59] P. Valent, “Mast cell activation syndromes: Definition and classification,” *Allergy*, vol. 68, no. 4, pp. 417–424, 2013. DOI: https://doi.org/10.1111/all.12126. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/all.12126. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.12126.

[60] S. C. Glover, M. C. Carter, P. Korošec, *et al.*, “Clinical relevance of inherited genetic differences in human tryptases: Hereditary alpha-tryptasemia and beyond,” *Annals of Allergy, Asthma & Immunology*, vol. 127, no. 6, pp. 638–647, 2021, ISSN: 1081-1206. DOI: https://doi.org/10.1016/j.anai.2021.08.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1081120621005676.

[61] M. Gangireddy, *Systemic mastocytosis*, Jul. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK544345/#:~:text=Mastocytosis%20is%20a%20constellation%20of,is%20known%20as%20systemic%20mastocytosis..

[62] [Online]. Available: https://www.aaaai.org/conditions-treatments/related-conditions/systemic-mastocytosis.

[63] [Online]. Available: https://www.nhs.uk/conditions/mastocytosis/.

[64] Apr. 2024. [Online]. Available: https://tmsforacure.org/overview/systemic-mastocytosis/.

[65] J. Abramson, J. Adler, J. Dunger, *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," en, *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w. [Online]. Available: https://www.nature.com/articles/s41586-024-07487-w (visited on 08/02/2024).

[66] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Medical Image Analysis*, vol. 20, no. 1, pp. 1–4, 2016.

[67] C. M. Eckhardt *et al.*, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 31, pp. 376–381, 2023.

[68] P. Rajpurkar *et al.*, "Ai in health and medicine," *Nature Medicine*, 2022.

[69] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, e271–e297, 2019.

[70] C. J. Kelly *et al.*, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019.

[71] J. G. Richens *et al.*, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020.

[72] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[73] K. Lu *et al.*, "Identification of novel biomarkers in hunner's interstitial cystitis using the cibersort, an algorithm based on machine learning," *BMC Urology*, vol. 21, no. 109, 2021. DOI: 10.1186/s12894-021-00875-8.

[74] S. Bohjanen *et al.*, "Machine learning model quantifies mast cells in biopsies of psoriatic lesions imaged with confocal microscopy," *Skin Research and Technology*, 2024. DOI: 10.1111/srt.13763.

[75] L. J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. L. Turukalo, *et al.*, "Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment," *Frontiers in Microbiology*, vol. 12, p. 634 511, 2021. DOI: 10.3389/fmicb.2021.634511.

[76] V. Fortino *et al.*, "Machine-learning–driven biomarker discovery for the identification and classification of oncological diseases," *Nature Biomedical Engineering*, 2020. DOI: 10.1038/s41598-023-45932-4.

[77] Z. Wang *et al.*, "Machine learning-based cluster analysis of immune responses in inflammatory diseases," *Scientific Reports*, 2023. DOI: 10.1038/s41598-023-45932-4.

[78] C. J. Kelly *et al.*, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019. DOI: 10.1186/s12916-019-1426-2.

[79] Z. et al., "Machine learning–based identification and characterization of genetic markers in multi-omics data," *Journal of Medical Genetics*, 2024. DOI: 10.1038/s41591-021-01614-0.

[80] A. Nguyen *et al.*, "The application of artificial intelligence to accelerate g protein-coupled receptor drug discovery," *British Journal of Pharmacology*, vol. 180, pp. 1234–1250, 2023. DOI: 10.1111/bph.15849.

[81] T. Ryan *et al.*, "Artificial intelligence and machine learning for clinical pharmacology," *British Journal of Clinical Pharmacology*, vol. 95, no. 5, pp. 1105–1120, 2023. DOI: 10.1111/bcp.15188.

[82] N. Terranova *et al.*, "Artificial intelligence for quantitative modeling in drug discovery and development," *Clinical Pharmacology and Therapeutics*, vol. 114, no. 1, pp. 54–68, 2023. DOI: 10.1002/cpt.2400.

[83] L. E. McCoubrey, M. Elbadawi, G. Orlu, S. Gaisford, and A. W. Basit, "Harnessing machine learning for development of microbiome therapeutics," *Gut Microbes*, vol. 13, no. 1, p. 1 872 323, 2021. DOI: 10.1080/19490976.2021.1872323.

[84] B. P. Society, "Integration of ai in the optimization of g protein-coupled receptor targeting drugs," *British Journal of Pharmacology*, vol. 180, pp. 1251–1265, 2023. DOI: 10.1111/bph.15850.

[85] J. Smith *et al.*, "Insights into artificial intelligence utilization in drug discovery," *Journal of Medicinal Chemistry*, vol. 66, no. 8, pp. 3599–3615, 2023. DOI: 10.1021/jm400104q.

[86] F. C. Udegbe *et al.*, "Machine learning in drug discovery: A critical review of applications and challenges," *Computer Science & IT Research Journal*, vol. 5, no. 4, pp. 892–902, 2024. DOI: 10.51594/csitrj.v5i4.1048.

[87] P. S. Foundation, *Welcome to Python.org*, en. [Online]. Available: https://www.python.org/doc/ (visited on 08/05/2024).

[88] T. Kluyver, B. Ragan-Kelley, F. Pérez, *et al.*, "Jupyter notebooks – a publishing format for reproducible computational workflows," *International Conference on Electronic Publishing*, pp. 87–90, Jan. 2016. DOI: 10.3233/978-1-61499-649-1-87. [Online]. Available: http://doi.org/10.3233/978-1-61499-649-1-87.

[89] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[90] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[91] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[92] A. Paszke, S. Gross, F. Massa, *et al.*, *Pytorch: An imperative style, high-performance deep learning library*, 2019. arXiv: 1912.01703 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1912.01703.

[93] M. Nakhwan and R. Duangsoithong, "Comparison analysis of data augmentation using bootstrap, gans and autoencoder," in *2022 14th International Conference on Knowledge and Smart Technology (KST)*, 2022, pp. 18–23. DOI: 10.1109/KST53302.2022.9729065.

[94] F. H. K. dos Santos Tanaka and C. Aranha, "Data augmentation using gans," *CoRR*, vol. abs/1904.09135, 2019. [Online]. Available: http://arxiv.org/abs/1904.09135.

[95] J. Jordon, L. Szpruch, F. Houssiau, *et al.*, *Synthetic Data – what, why and how?* en, arXiv:2205.03257 [cs], May 2022. [Online]. Available: http://arxiv.org/abs/2205.03257 (visited on 08/07/2024).

[96] J. Brownlee, *Train Neural Networks With Noise to Reduce Overfitting*, en-US, Dec. 2018. (visited on 08/08/2024).

[97] M. S. Somanna, *Guide to Adding Noise to your Data using Python and Numpy*, en, Aug. 2023. [Online]. Available: https://medium.com/@ms_somanna/guide-to-adding-noise-to-your-data-using-python-and-numpy-c8be815df524 (visited on 08/07/2024).

[98] H. Lobbes, Q. Reynaud, S. Mainbourg, J. C. Lega, I. Durieu, and S. Durupt, "[tryptase: A practical guide for the physician]," *La Revue De Medecine Interne*, vol. 41, no. 11, pp. 748–755, Nov. 2020, ISSN: 1768-3122. DOI: 10.1016/j.revmed.2020.06.006.

[99] imbalanced-learn developers. "Smote." Accessed: 2024-08-10. (2024), [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

[100] V. G. Biju, A.-M. Schmitt, and B. Engelmann, "Assessing the influence of sensor-induced noise on machine-learning-based changeover detection in cnc machines," *Sensors*, vol. 24, no. 2, p. 330, 2024, Submission received: 3 November 2023 / Revised: 18 December 2023 / Accepted: 3 January 2024 / Published: 5 January 2024. DOI: 10.3390/s24020330.

[101] K. Wang, V. Muthukumar, and C. Thrampoulidis, "Benign overfitting in multiclass classification: All roads lead to interpolation," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=0O5jpovbdHO.

[102] *Human microbiome project data analysis and coordination center*, Accessed: 2024-08-03, 2024. [Online]. Available: https://www.hmpdacc.org/.

[103] L. Zhu, X. Jian, B. Zhou, *et al.*, "Gut microbiota facilitate chronic spontaneous urticaria," *Nature Communications*, vol. 15, p. 112, 2024. DOI: 10.1038/s41467-023-44373-x.

[104] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano, "Synthetic data generation: State of the art in health care domain," *Computer Science Review*, vol. 48, p. 100 546, May 2023, ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2023.100546. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013723000138 (visited on 08/09/2024).

[105] P. Li, H. Luo, B. Ji, *et al.*, "Machine learning for data integration in human gut microbiome," *Microb Cell Fact*, vol. 21, p. 241, 2022. DOI: 10.1186/s12934-022-01973-4.

[106] L. Yang and J. Chen, "A comprehensive evaluation of microbial differential abundance analysis methods: Current status and potential solutions," *Microbiome*, vol. 10, p. 130, 2022. DOI: 10.1186/s40168-022-01320-0. [Online]. Available: https://doi.org/10.1186/s40168-022-01320-0.

[107] S. Weiss, Z. Z. Xu, S. Peddada, and et al., "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, p. 27, 2017. DOI: 10.1186/s40168-017-0237-y. [Online]. Available: https://doi.org/10.1186/s40168-017-0237-y.

[108] S. W. Kembel, M. Wu, J. A. Eisen, and J. L. Green, "Incorporating 16s gene copy number information improves estimates of microbial diversity and abundance," *PLOS Computational Biology*, 2012. DOI: 10.1371/journal.pcbi.1002743. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002743.

[109] A. L. M. Levada, "Pca-kl: A parametric dimensionality reduction approach for unsupervised metric learning," *Advances in Data Analysis and Classification*, vol. 15, pp. 829–868, 2021. DOI: 10.1007/s11634-020-00434-3. [Online]. Available: https://doi.org/10.1007/s11634-020-00434-3.

[110] Y. Shi, L. Zhang, C. B. Peterson, and et al., "Performance determinants of unsupervised clustering methods for microbiome data," *Microbiome*, vol. 10, no. 25, 2022. DOI: 10.1186/s40168-021-01199-3. [Online]. Available: https://doi.org/10.1186/s40168-021-01199-3.

[111] A. Behrouzi, A. Nafari, and S. Siadat, "The significance of microbiome in personalized medicine," *Clinical and Translational Medicine*, vol. 8, no. 16, 2019. DOI: 10.1186/s40169-019-0232-y. [Online]. Available: https://doi.org/10.1186/s40169-019-0232-y.

[112] K. Ratiner, D. Ciocan, S. K. Abdeen, and E. Elinav, "Utilization of the microbiome in personalized medicine," *Nature Reviews Microbiology*, 2023. DOI: 10.1038/s41579-023-00998-9. [Online]. Available: https://doi.org/10.1038/s41579-023-00998-9.

[113] M. Mittermaier *et al.*, "Bias in ai-based models for medical applications: Challenges and mitigation strategies," *NPJ Digital Medicine*, vol. 6, no. 1, p. 113, 2023. DOI: 10.1038/s41746-023-00858-z.

Appendix

A Code Operation Manual

Set Up

Before you begin, ensure that you have the following prerequisites installed on your system:

- Python 3.7+: The codebase is developed in Python, and it is recommended to use Python version 3.7 or higher.
- Git: Required to clone the repository.
- Virtual Environment: Strongly recommended for managing project dependencies.

Step 1: install the required dependencies:

```
pip install pandas numpy matplotlib seaborn scikit-learn joblib xgboost
tensorflow keras scipy ipython opencv-python pillow dask[bag] ultralytics
```

Step 2: Clone the project repository from GitHub to your local machine using the following commands:

```
git clone https://github.com/username/thesis-project.git
cd thesis-project
```

File Directory

The project is organized into several directories, each serving a specific purpose. Table 2 outlines the important files in our repository.

Running the Code Base

To run the project and review the analysis:

1. Open the relevant Jupyter notebooks or scripts and verify that the data paths are correct.
2. Execute the notebooks to preprocess data, train models, and generate results.
3. Use the ‘script.py‘ for running predictions with pre-trained models.

Please note that the project cannot be executed without the data files. We did not attach the data files in our submitted code base for size and medical privacy reasons. Please reach out if you need further assistance or would like to receive the data in a secure way.

Table 2: *List of Artifact Files for this Thesis*

File	Directory	Description
Image Classification.ipynb	Image Data/Image Classification	Contains notebooks related to image classification tasks.
ObjectDetection.ipynb	Image Data/Object Detection	Includes notebooks for object detection tasks involving histopathology images.
MicrobiomeDataAug.ipynb	Microbiome Data/	Notebook for data augmentation specific to microbiome data.
MicrobiomeModels.ipynb	Microbiome Data/	Notebook for modeling and analyzing microbiome data.
Small_DataSet_ Exploration.ipynb	Tabular Data/Small Dataset	Notebook for initial data exploration.
Small_DataSet_ Model.ipynb	Tabular Data/Small Dataset	Notebook for building and evaluating models on the small dataset.
TraditionalAug_ ModelTraining.ipynb	Tabular Data/Tabular Data Augmenta- tion/Model Training	Notebook to train on augmented data.
TraditionalAugmentation_ Tabular.ipynb	Tabular Data/Tabular Data Augmentation	Notebook for traditional data augmentation techniques.
GAN_DataAugmentation_ Tabular.ipynb	Tabular Data/Tabular Data Augmentation	Notebook for applying GAN-based augmentation.
utils_ApplyPipeline.py	Utils Scripts/	Script to apply the data processing pipeline to new data.
utils_DataExploration.py	Utils Scripts/	Functions for exploratory data analysis.
utils_diagnosisgroups.py	Utils Scripts/	Functions for grouping data based on diagnosis categories.
utils_preprocessing.py	Utils Scripts/	Data preprocessing utilities.
script.py	Models and Scripts/	A script that ties together various parts of the codebase for executing a full analysis pipeline.

B Testing Summary

In this project, we tested our machine learning models using a traditional approach where 20% of the data was held out as a testing set to evaluate model performance.

We focused on ensuring that our models were not only accurate but also generalized well. The evaluation metrics included accuracy, precision, recall, and F1-score providing insights into how well the models performed in classifying and predicting outcomes based on both the original and augmented data. For the microbiome's unsupervised model, the model was tested manually by comparing the cluster analysis output to the bacterial I manipulated during the augmentation phase.

Throughout the testing phase, we monitored the models' performance and made necessary adjustments to optimize results. While the initial outcomes were promising, indicating that the models could effectively handle the data, continuous fine-tuning was essential to address any discrepancies or biases observed during testing. This straightforward approach was critical for validating the models' applicability to real-world scenarios, though further testing with more diverse datasets would be beneficial to confirm these findings.

C Bioinformatics & 16s Sequencing

Bioinformatics is essential in microbiome research. It combines biology and computational techniques to analyze data from microorganisms. There are two important methods for sequencing microbiome data: shotgun and 16s. The human microbiome project data was created using 16s, therefore, the following explanation will focus on the 16s bioinformatics pipeline. 16S rRNA sequencing allows scientists to identify and quantify bacterial species in a sample by analyzing the 16S ribosomal RNA gene found in all bacteria and archaea and comparing the sequences to a database of known bacterial genes. The wisdom gained from 16S rRNA sequencing reveal the diversity and composition of microbial communities, offering valuable perspectives for research in health, disease, and environmental microbiomes. Please see Appendix J for a more detailed process flow of this pipeline.

D Ethics Approval

University of St Andrews | FOUNDED 1413

School of Computer Science Ethics Committee

28 June 2024

Dear Lena and Lea,

Thank you for submitting your ethical application which was considered by the School Ethics Committee.

The School of Computer Science Ethics Committee, acting on behalf of the University Teaching and Research Ethics Committee (UTREC), has approved this application:

Approval Code:	CS18003	**Approved on:**	28.06.24	**Approval Expiry:**	28.06.29
Project Title:	Machine Learning for Histopathology in Mast Cell Diseases				
Researcher(s):	Lena Korfmacher and Lea Brody-heine				
Supervisor(s):	Professor Tom Kelsey				

The following supporting documents are also acknowledged and approved:

1. Application Form

Approval is awarded for 5 years, see the approval expiry data above.

If your project has not commenced within 2 years of approval, you must submit a new and updated ethical application to your School Ethics Committee.

If you are unable to complete your research by the approval expiry date you must request an extension to the approval period. You can write to your School Ethics Committee who may grant a discretionary extension of up to 6 months. For longer extensions, or for any other changes, you must submit an ethical amendment application.

You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
  - the details provided in your ethical application
  - the University's Principles of Good Research Conduct
  - the conditions of any funding associated with your work
- that you obtain all applicable additional documents (see the 'additional documents' webpage for guidance) before research commences.

You should retain this approval letter with your study paperwork.

Yours sincerely,

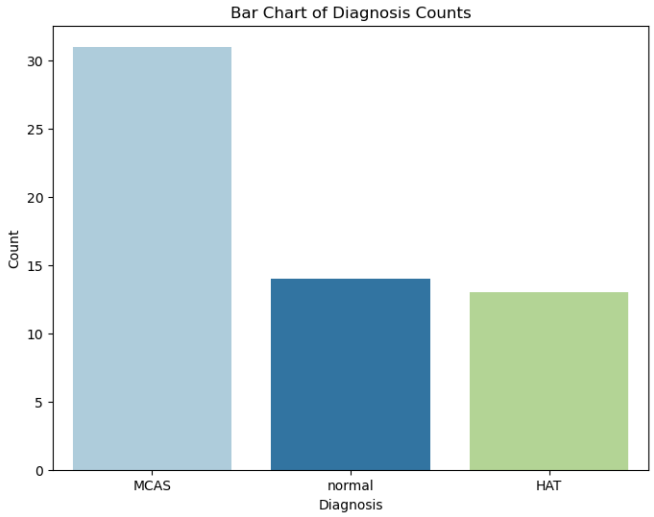
Wendy Boyter

SEC Administrator

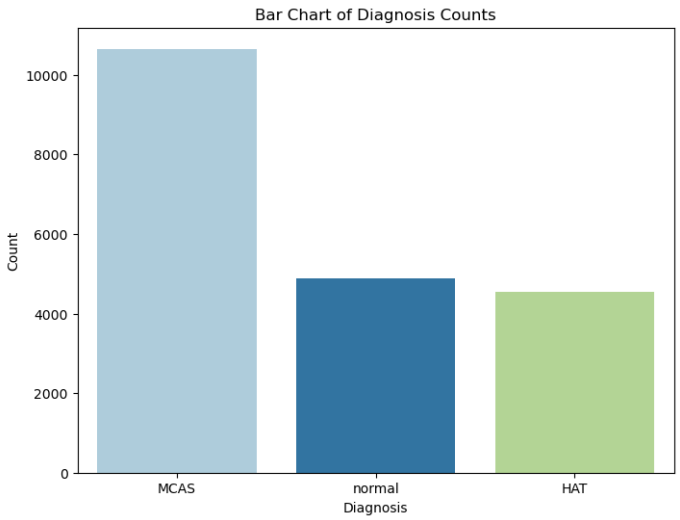
School of Computer Science Ethics Committee
Dr Olexandr Konovalov/Convenor, Jack Cole Building, North Haugh, St Andrews, Fife, KY16 9SX
Telephone: 01334 463273 Email: ethics-cs@st-andrews.ac.uk

E Tabular Data EDA Figures

Figure 15: *Imbalanced Classes*



(a) *Original Data*



(b) *Augmented Data*

Figure 16: *Original Data Box Plot Outliers*

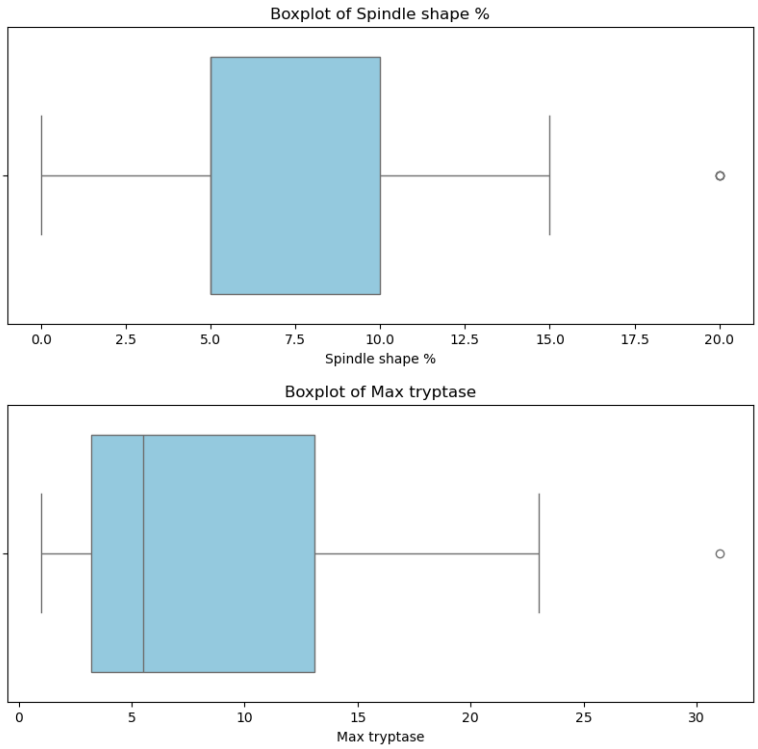


Figure 17: *Augmented Data Box Plot Outliers*

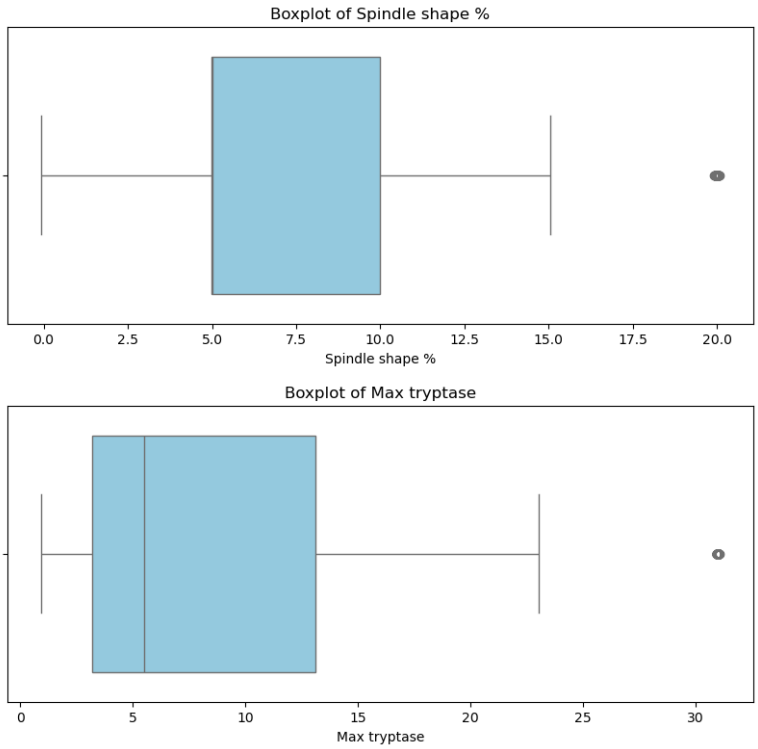
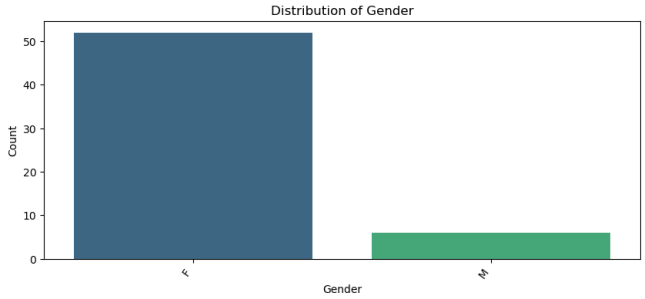
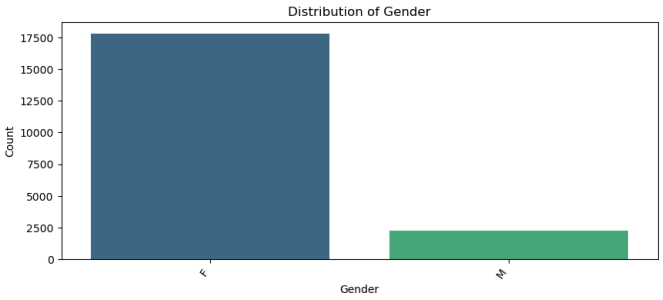


Figure 18: *Gender Distribution Comparison*

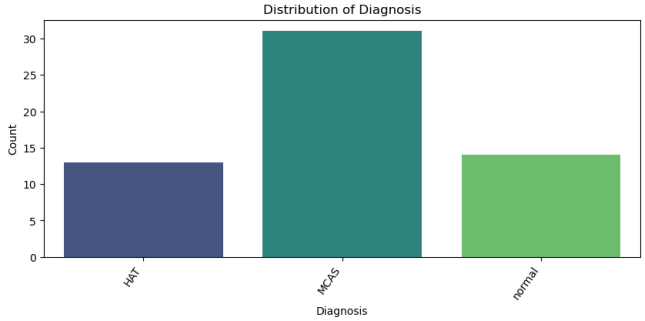


(a) *Original Data*

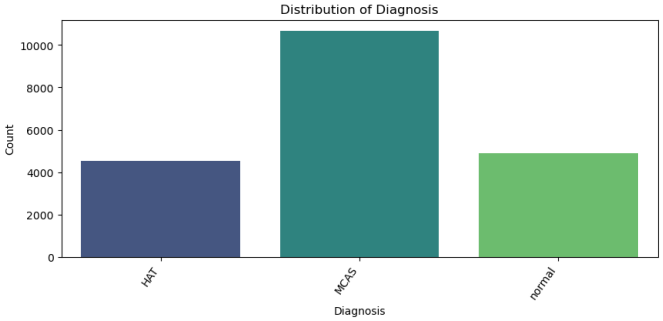


(b) *Augmented Data*

Figure 19: *Diagnosis Distribution Comparison*

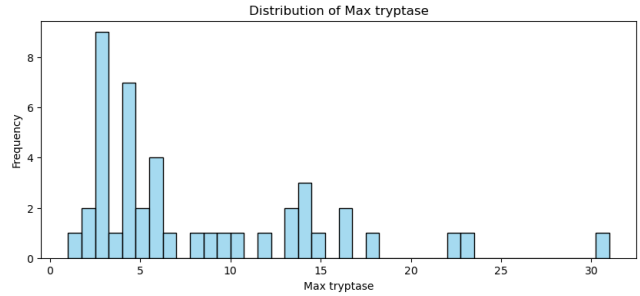


(a) *Original Data*

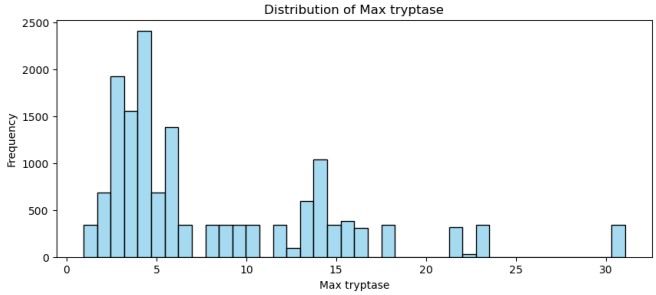


(b) *Augmented Data*

Figure 20: *Comparison of Max Tryptase Levels Distribution*

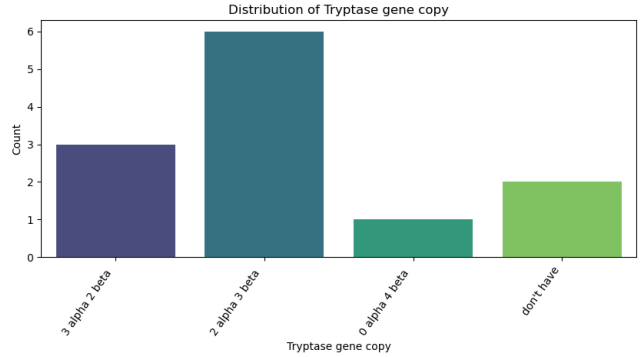


(a) *Original Data*

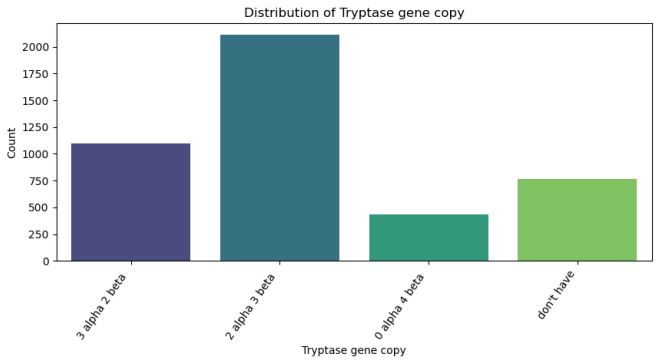


(b) *Augmented Data*

Figure 21: *Tryptase Gene Copy Number Distribution Comparison*

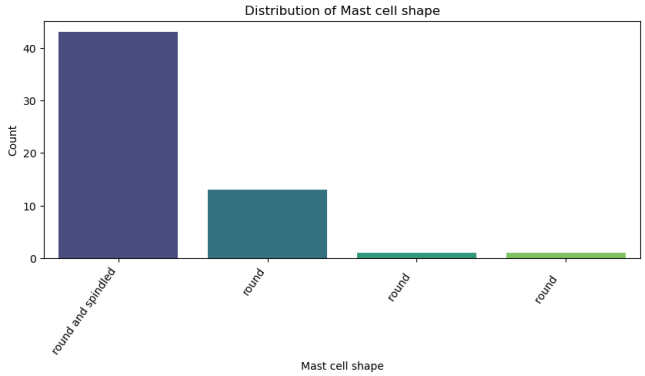


(a) *Original Data*

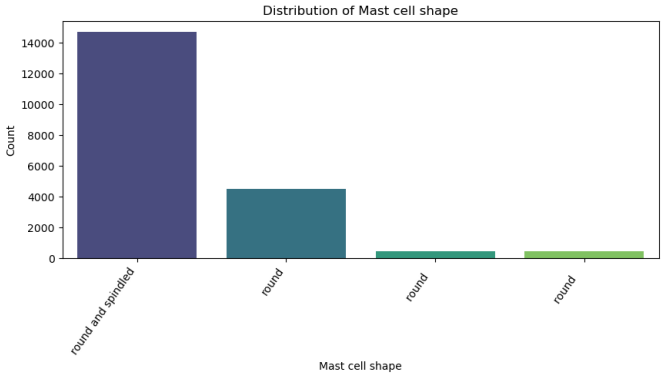


(b) *Augmented Data*

Figure 22: *Distribution of Mast Cell Shapes Comparison*

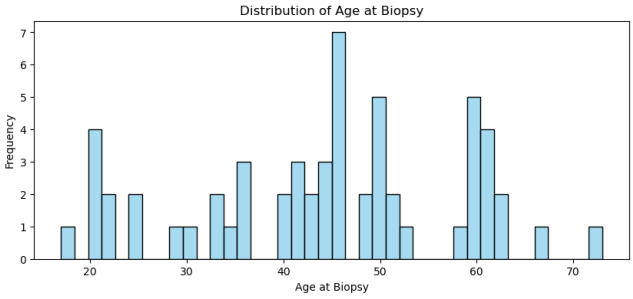


(a) *Original Data*

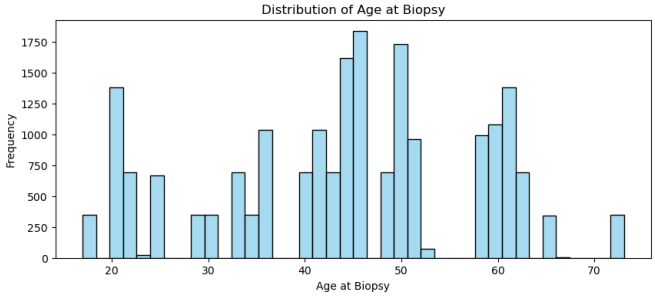


(b) *Augmented Data*

Figure 23: *Age Distribution at Biopsy in Original vs. Augmented Data*

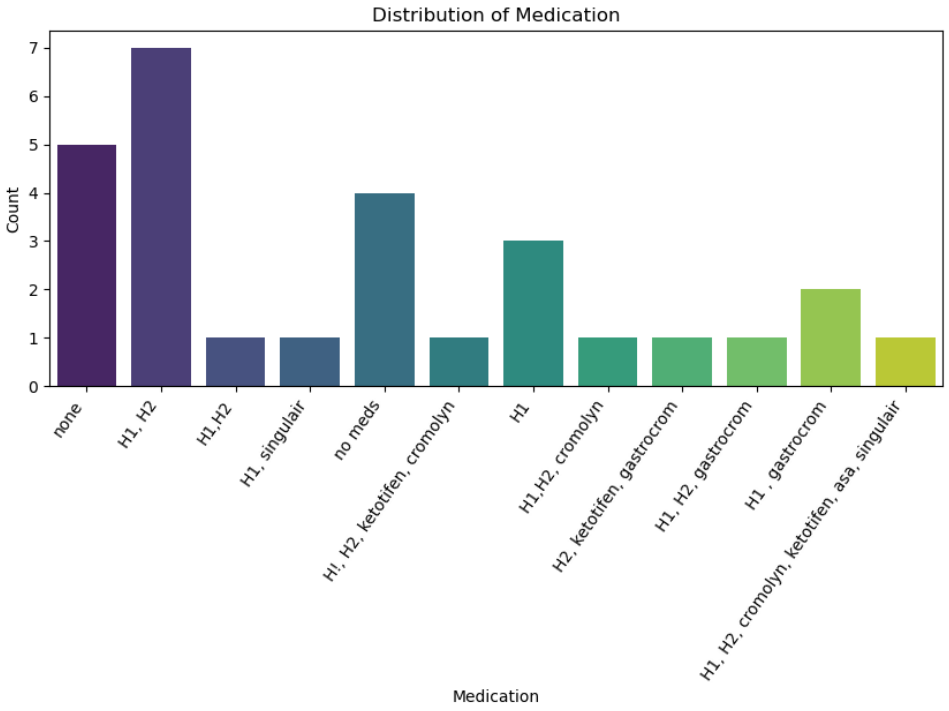


(a) *Original Data*

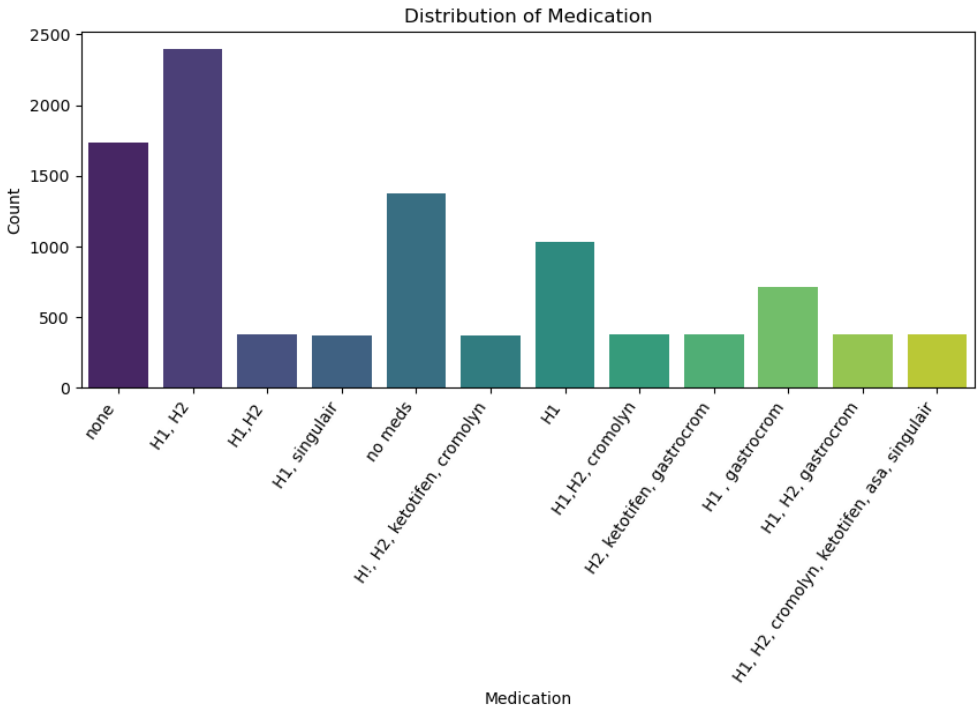


(b) *Augmented Data*

Figure 24: *Medication Use Distribution in Original vs. Augmented Data*

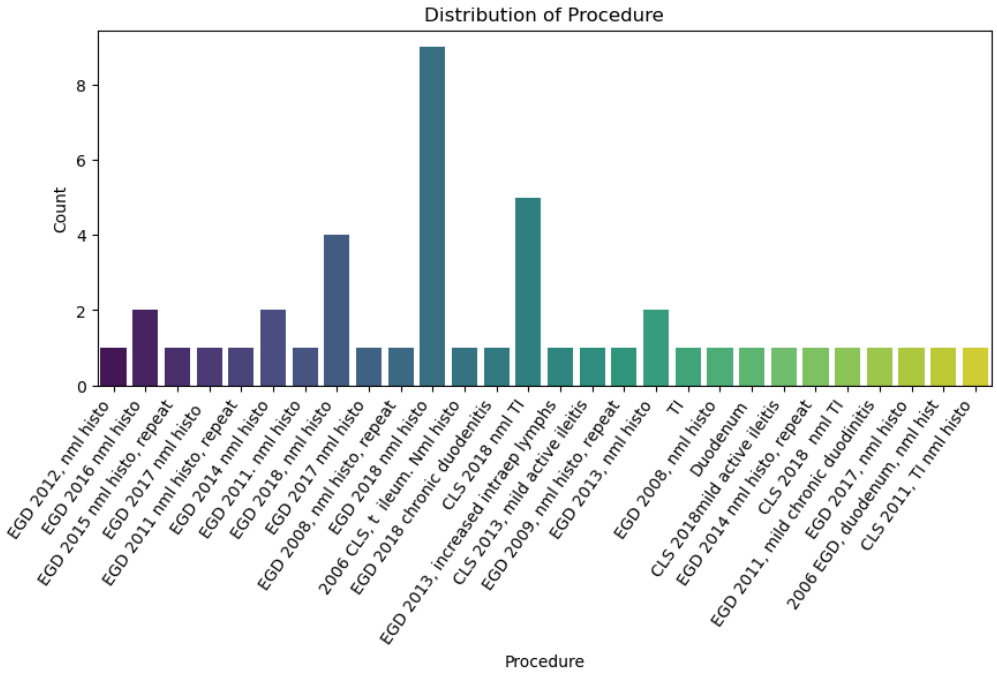


(a) *Original Data*

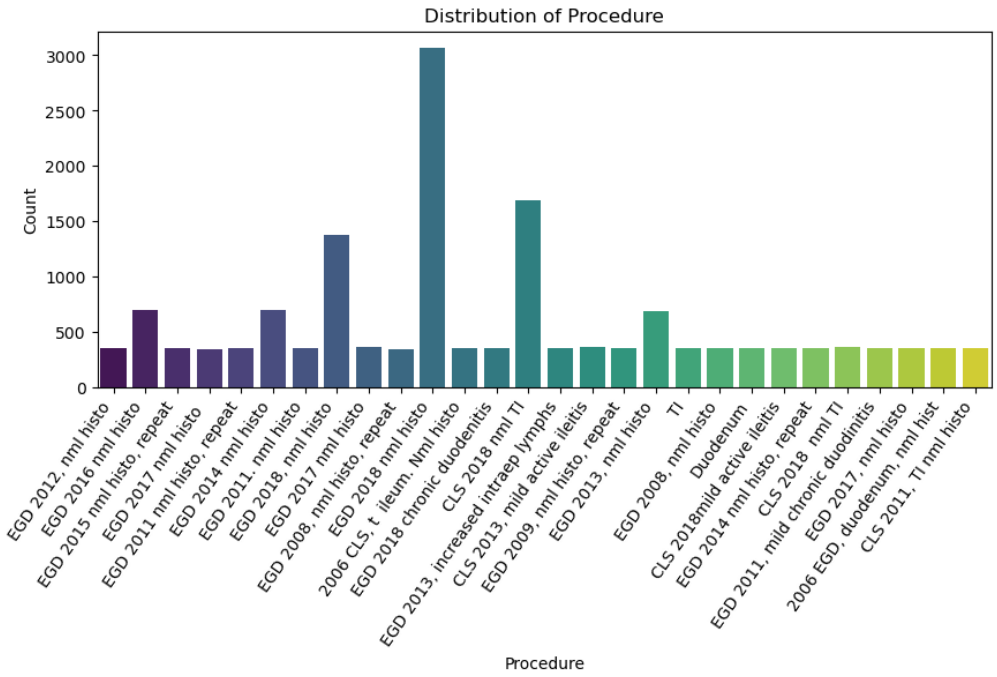


(b) *Augmented Data*

Figure 25: *Procedure Distribution Comparison*

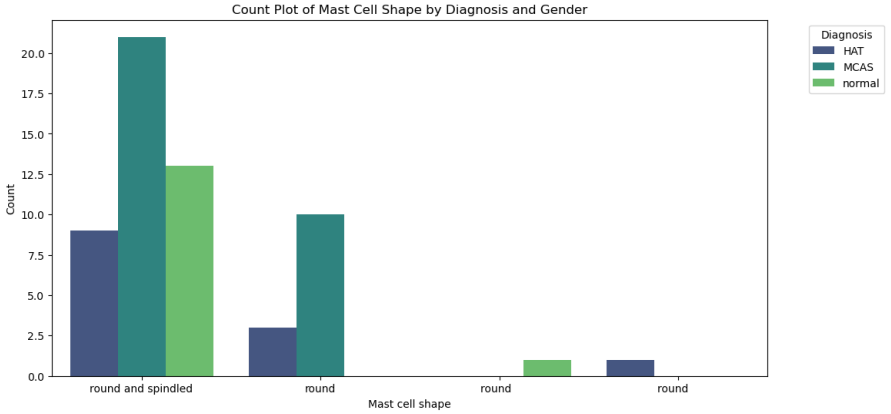


(a) *Original Data*

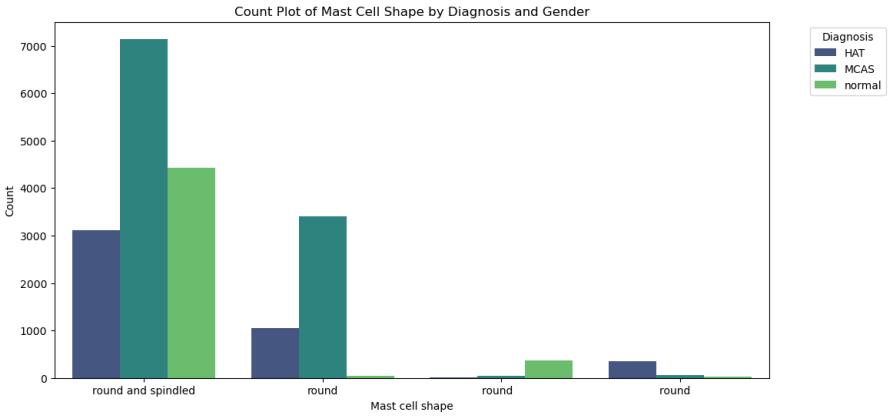


(b) *Augmented Data*

Figure 26: *Comparison of Mast Cell Shape, Gender, and Diagnosis*

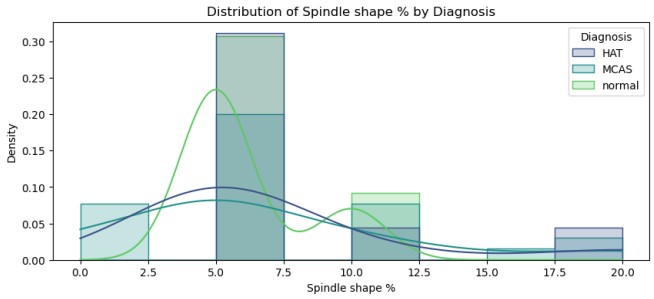


(a) *Original Data*

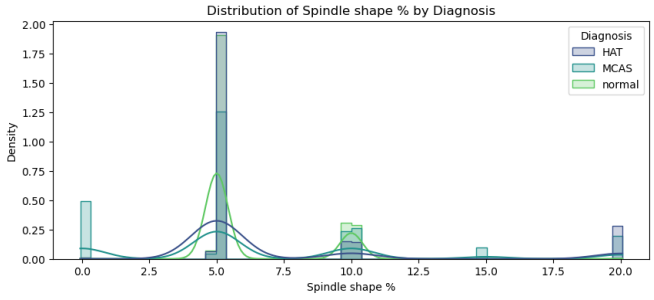


(b) *Augmented Data*

Figure 27: *Distribution of Mast Cell Shape by Diagnosis*

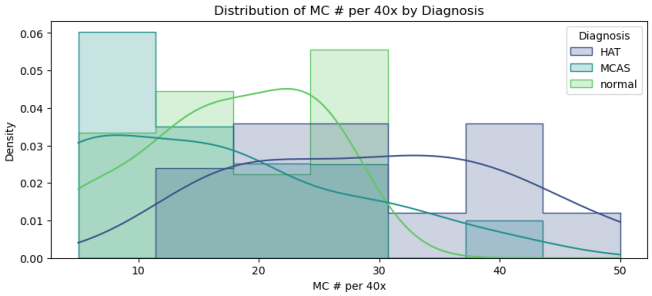


(a) *Original Data*

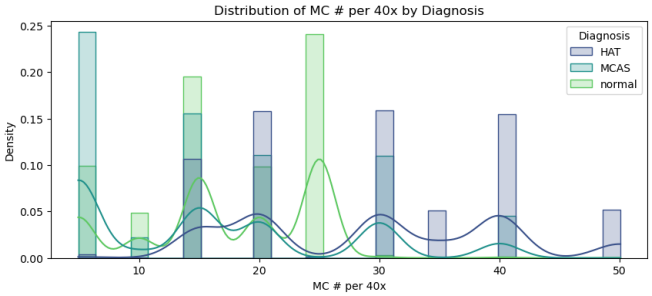


(b) *Augmented Data*

Figure 28: *Comparison of Mast Cell Count by Diagnosis*

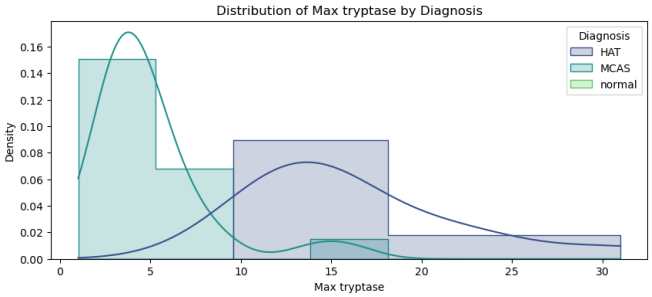


(a) *Original Data*

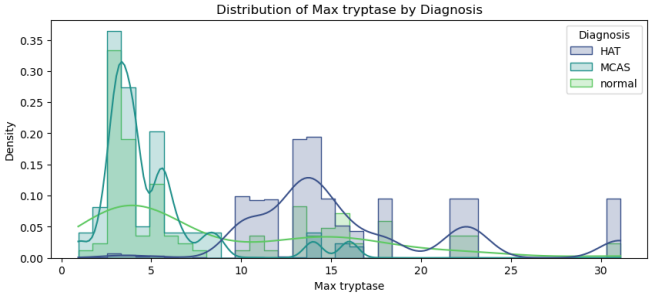


(b) *Augmented Data*

Figure 29: *Comparison of Maximum Tryptase Levels by Diagnosis*

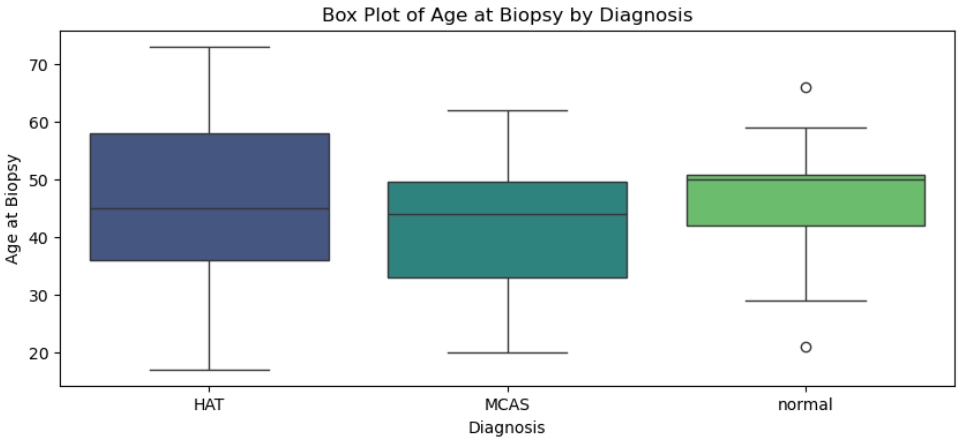


(a) *Original Data*

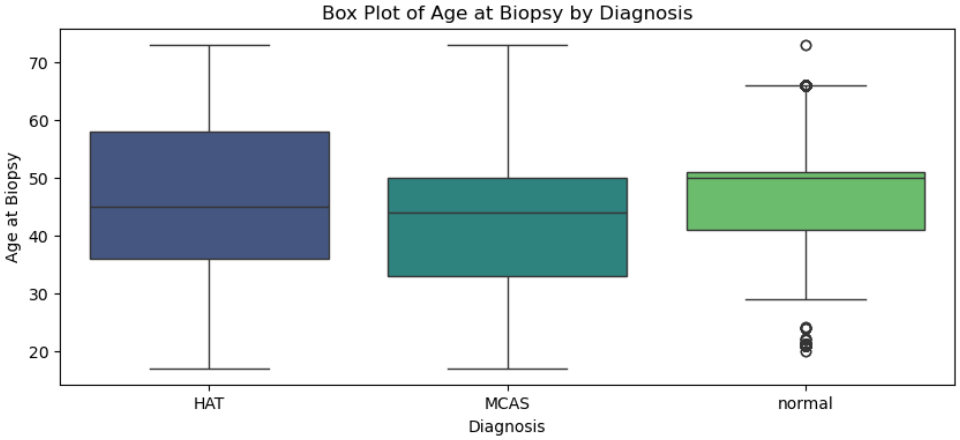


(b) *Augmented Data*

Figure 30: *Age Distribution by Diagnosis Comparison*



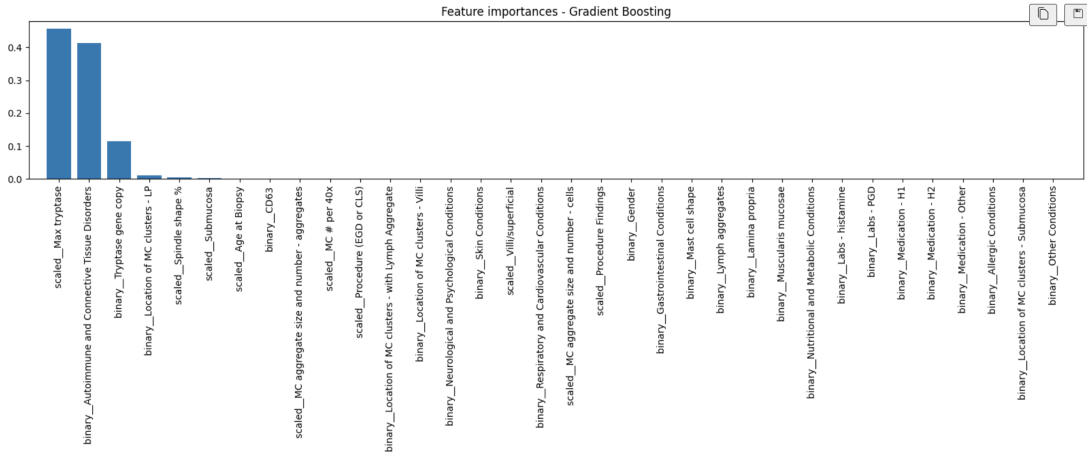
(a) *Original Data*



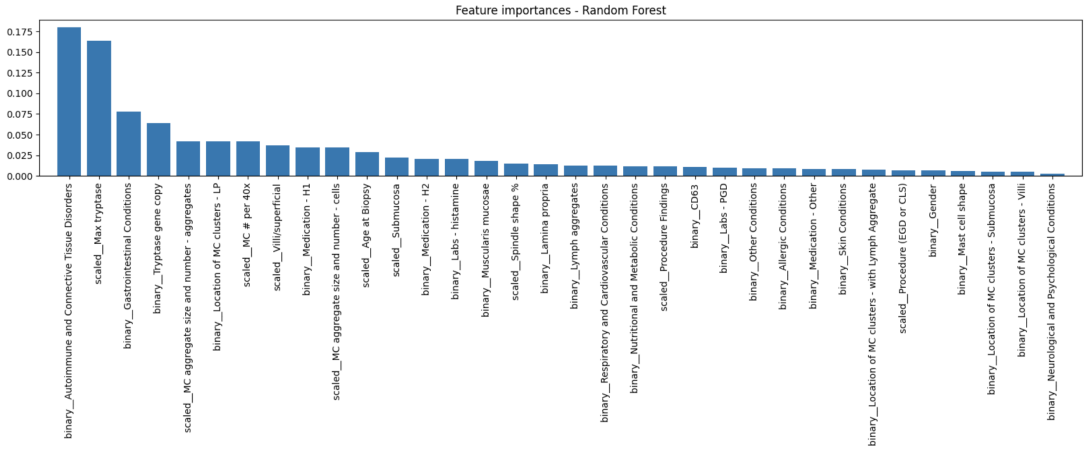
(b) *Augmented Data*

F Feature Importance

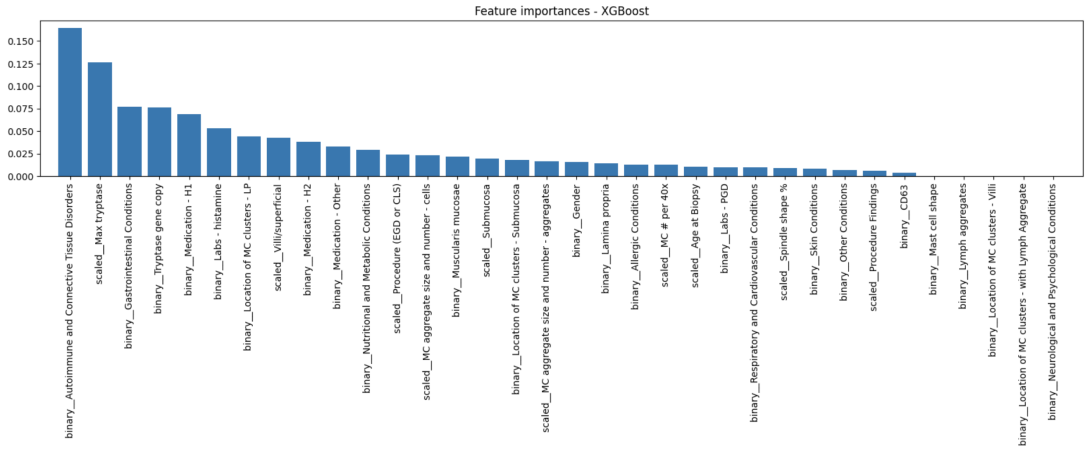
Figure 31: *Comparison of feature importances for selected models*



(a) *Feature importance of Gradient Boosting*



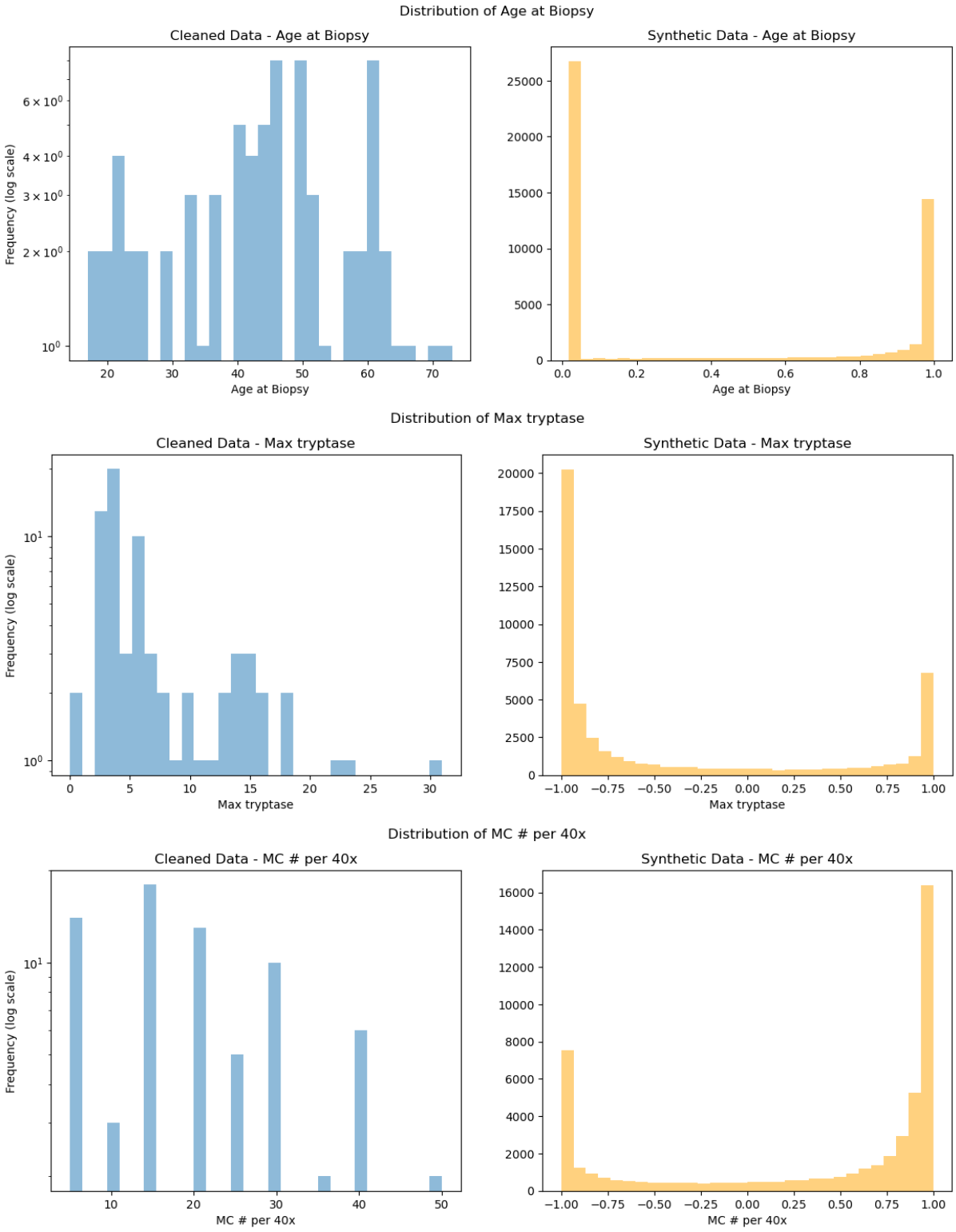
(b) *Feature importance of Random Forest*



(c) *Feature importdance of XGBoost*

G Tabular Data Augmentation Figures

Figure 32: *Comparison of features: Original & GAN Generated Data*



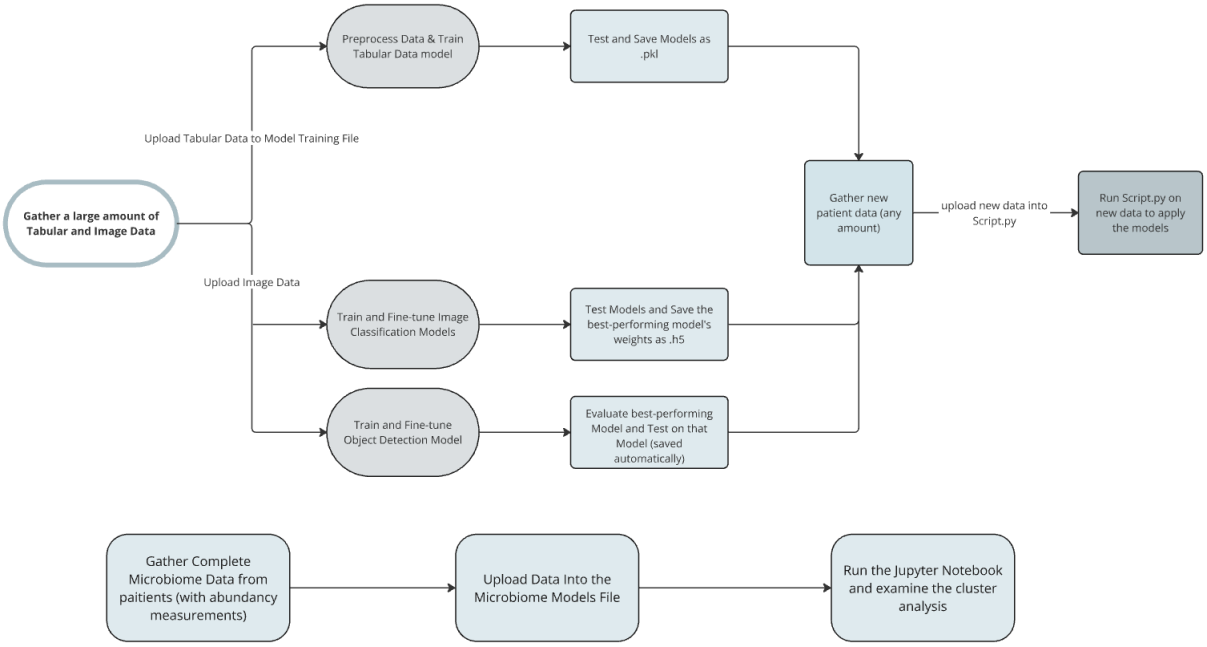
H Tabular Data Features

Table 3: *Final Features in the Tabular Dataset*

Feature	Description
Age at Biopsy	Age of patient at the time of the biopsy (numeric)
Gender	Gender of patient, female/male (binary)
Max tryptase	Maximum tryptase serum value (numeric)
Tryptase gene copy	Presence of excess tryptase gene copies (binary)
MC # per 40x	Mast cell density at 40x magnification (numeric)
Mast cell shape	Shape of mast cells, round/round and spindled (binary)
Lymph aggregates	Presence of lymph aggregates (binary)
Spindle shape %	Percentage of mast cell spindle shape (numeric)
Villi/superficial	Presence of mast cells in the villi/superficial (binary)
Lamina propria	Presence of mast cells in the lamina propria (binary)
Muscularis mucosae	Presence of mast cells in the muscularis mucosae (binary)
Submucosa	Presence of mast cells in the submucosa (binary)
CD63	CD63 staining, positive/negative (binary)
MC aggregate size and number - cells	Number of cells per mast cell aggregate (numeric)
MC aggregate size and number - aggregates	Number of mast cell aggregates (numeric)
Procedure (EGD or CLS)	Performed procedure, EGD/CLS (binary)
Procedure Findings	Procedure findings yes or no (binary)
Labs - histamine	Elevated histamine values yes or no (binary)
Labs - PGD	Elevated PGD values yes or no (binary)
Medication - H1	Patient takes H1 medication (binary)
Medication - H2	Patient takes H2 medication (binary)
Medication - Other	Patient takes other medication (binary)
Location of MC clusters - LP	Mast cell clusters in lamina propria (binary)
Location of MC clusters - Submucosa	Mast cell clusters in submucosa (binary)
Location of MC clusters - Villi	Mast cell clusters in villi (binary)
Location of MC clusters - with Lymph Aggregate	Mast cell clusters with lymph aggregates (binary)
Allergic Conditions	Patient has reported/diagnosed allergic conditions (binary)
Gastrointestinal Conditions	Patient has reported/diagnosed gastrointestinal conditions (binary)
Autoimmune and Connective Tissue Disorders	Patient has reported/diagnosed autoimmune or connective tissue disorders (binary)
Respiratory and Cardiovascular Conditions	Patient has reported/diagnosed respiratory or cardiovascular conditions (binary)
Skin Conditions	Patient has reported/diagnosed skin conditions (binary)
Neurological and Psychological Conditions	Patient has reported/diagnosed neurological or psychological conditions (binary)
Nutritional and Metabolic Conditions	Patient has reported/diagnosed nutritional or metabolic conditions (binary)
Other Conditions	Patient has reported/diagnosed other conditions (binary)

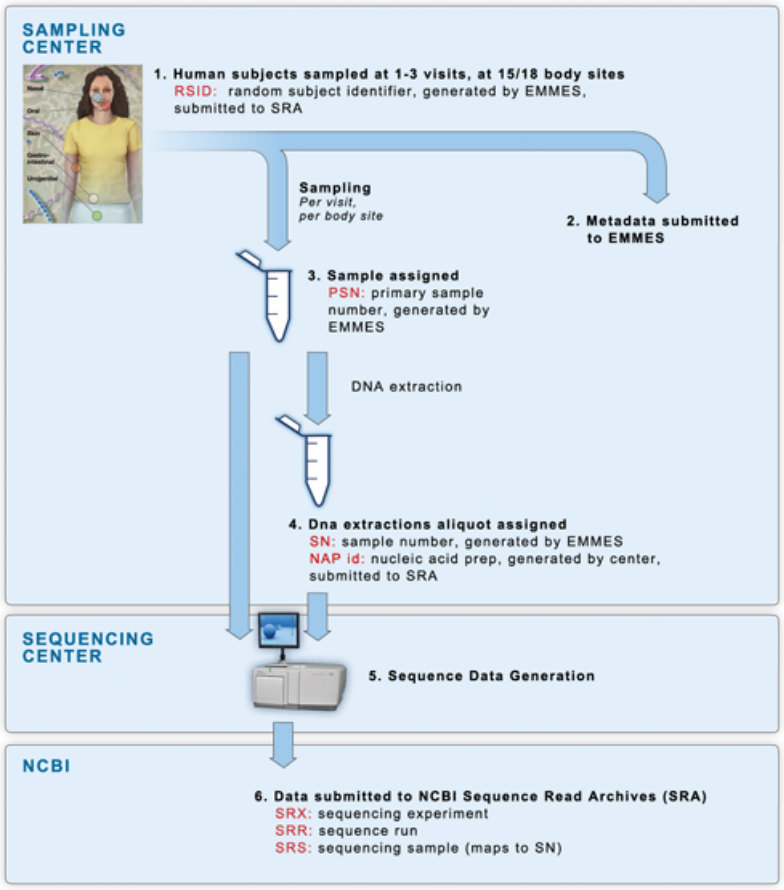
I Methodology Flow Chart

Figure 33: *Methodology Flow Chart*



J 16s Bioinformatics Pipeline

Figure 34: *16s Bioinformatics Pipeline*



Source: [102]

K Statistical Testing

Figure 35: *Results of T-Testing*

Significant differences found in the following bacteria:

	Bacterium	T-Statistic	P-Value	Significant
0	Klebsiella pneumoniae subsp. rhinoscleromatis	-inf	0.0	True
1	Klebsiella pneumoniae subsp. pneumoniae WGLW1	-inf	0.0	True
2	Klebsiella pneumoniae subsp. pneumoniae WGLW2	-inf	0.0	True
3	Klebsiella pneumoniae subsp. pneumoniae WGLW3	-inf	0.0	True
4	Klebsiella pneumoniae subsp. pneumoniae WGLW5	-inf	0.0	True
5	Klebsiella pneumoniae 909957	-inf	0.0	True
7	Klebsiella pneumoniae MJR8396D	-inf	0.0	True
15	Clostridium leptum DSM 753	1.377105e+18	0.0	True
16	Ruminococcus obeum ATCC 29174	inf	0.0	True

The t-test results reveal significant differences between the normal and i-MCAS groups for certain bacteria. Klebsiella pneumoniae strains and Ruminococcus obeum showed substantial variations, highlighting their potential as biomarkers for i-MCAS. The extreme T-Statistic values indicate clear differentiation between the groups, supporting the validity of the synthetic data and the pipeline's effectiveness.

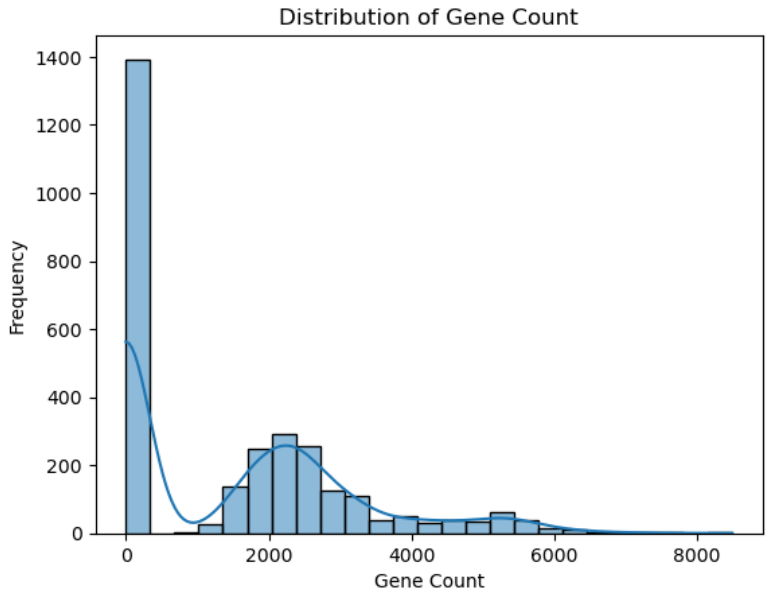
Figure 36: *Kolmogorov-Smirnov Testing Results*

```
Age at Biopsy - KS Statistic: 1.0, P-Value: 0.0
Max tryptase - KS Statistic: 0.9863013698630136, P-Value: 1.8987212322377924e-136
Tryptase gene copy - KS Statistic: 0.5042564383561644, P-Value: 1.037017950745205e-17
MC # per 40x - KS Statistic: 1.0, P-Value: 0.0
```

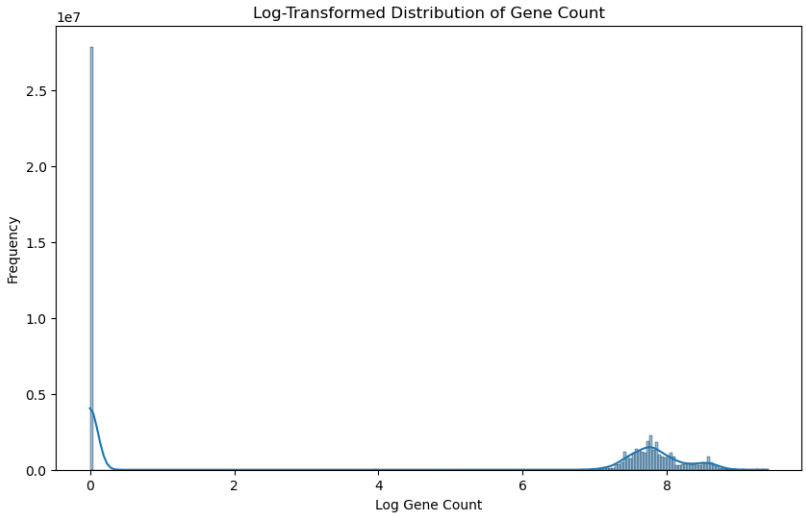
The test results show significant differences between the synthetic and original data distributions for key features such as 'Age at Biopsy' and 'Max tryptase'. The high KS statistics and low p-values indicate that the synthetic data distribution does not closely match the original data distribution, suggesting areas where the augmentation process may need refinement.

L Microbiome Original & Augmented Data EDA Figures

Figure 37: *Gene Count Distribution*



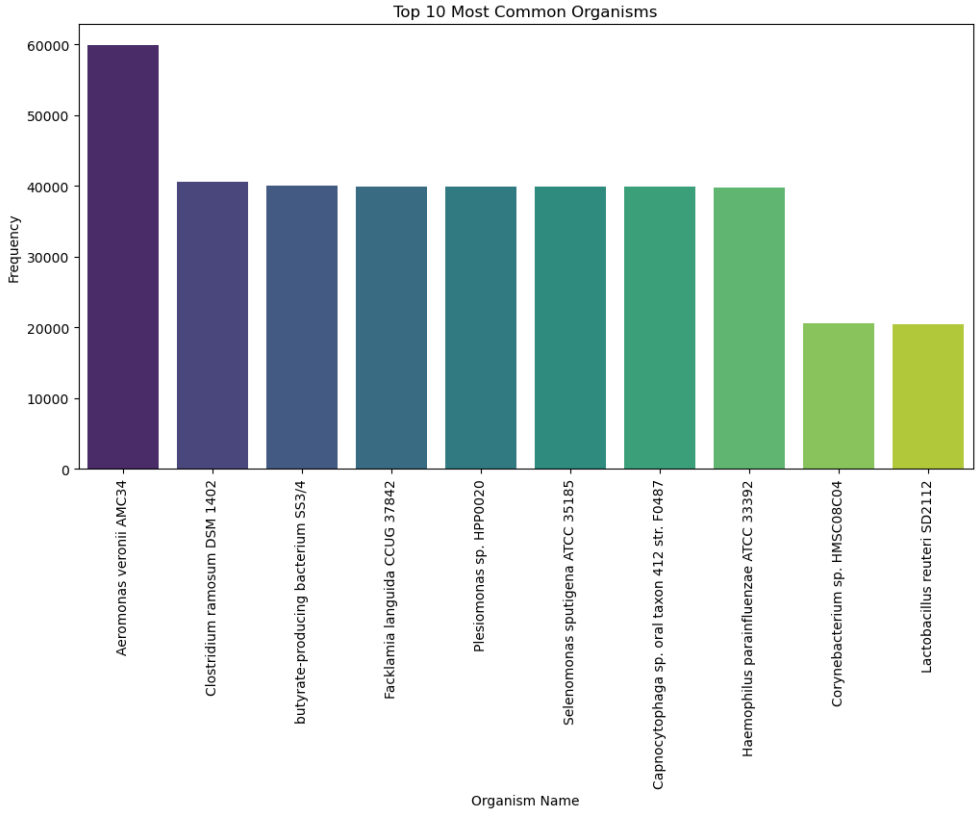
(a) *Original Data*



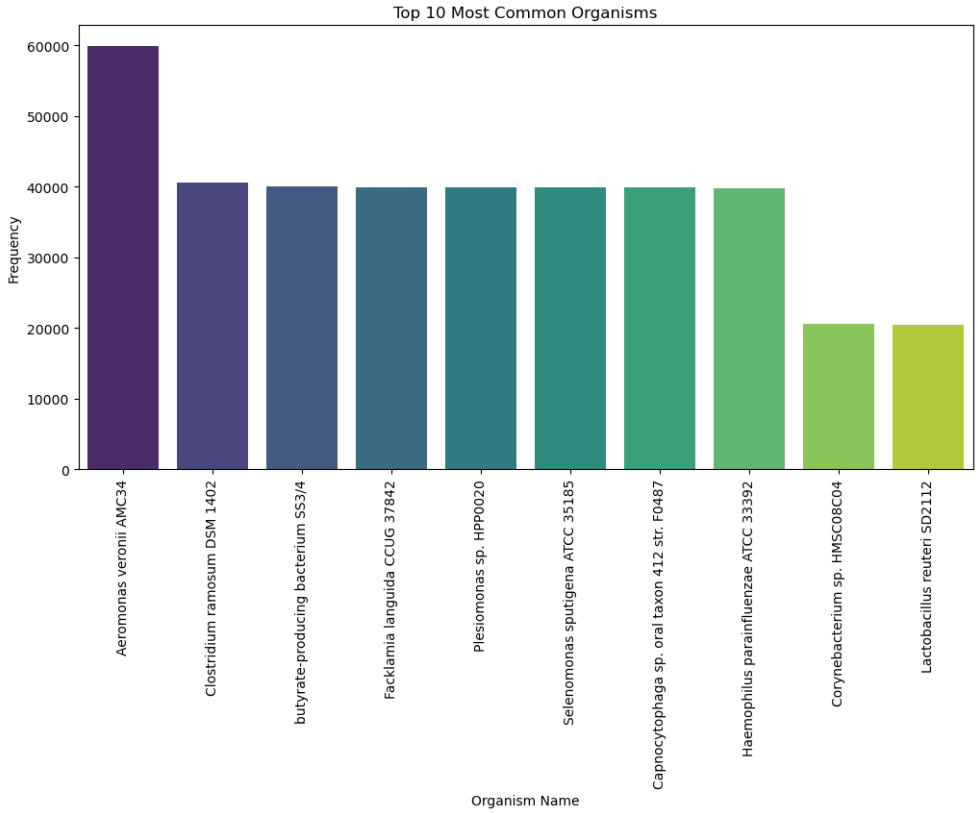
(b) *Augmented Data (log)*

The original data shows a right-skewed distribution of gene counts, while the augmented data uses a logarithmic transformation to compress the range and better visualize the differences in gene count distribution, especially among lower values.

Figure 38: *Top 10 Most Common Organism*

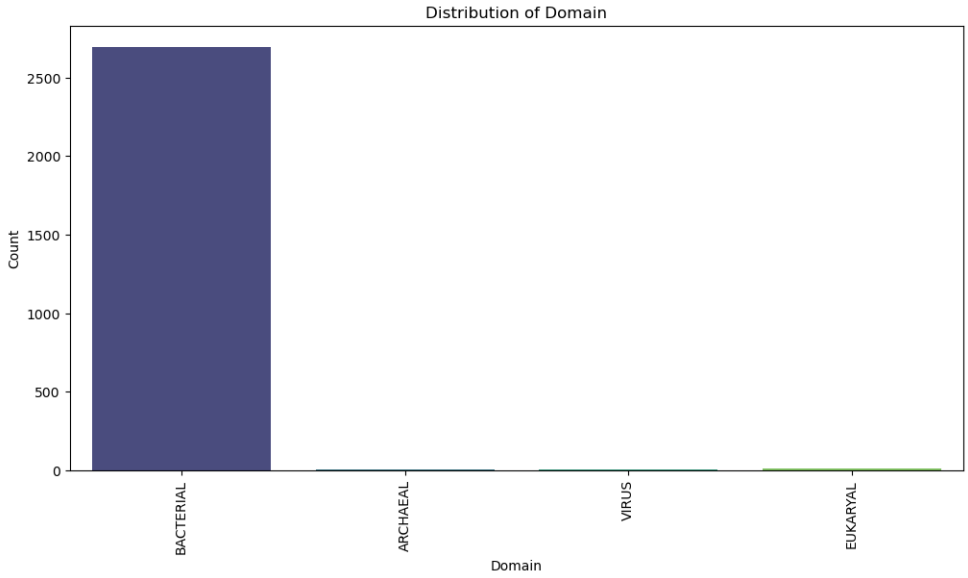


(a) *Original Data*

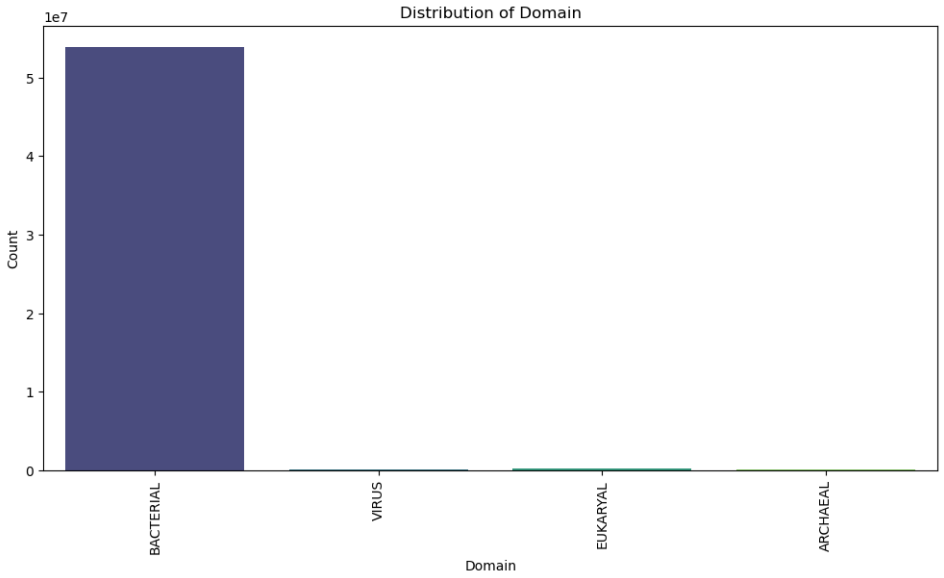


(b) *Augmented Data*

Figure 39: *Domain Distribution*

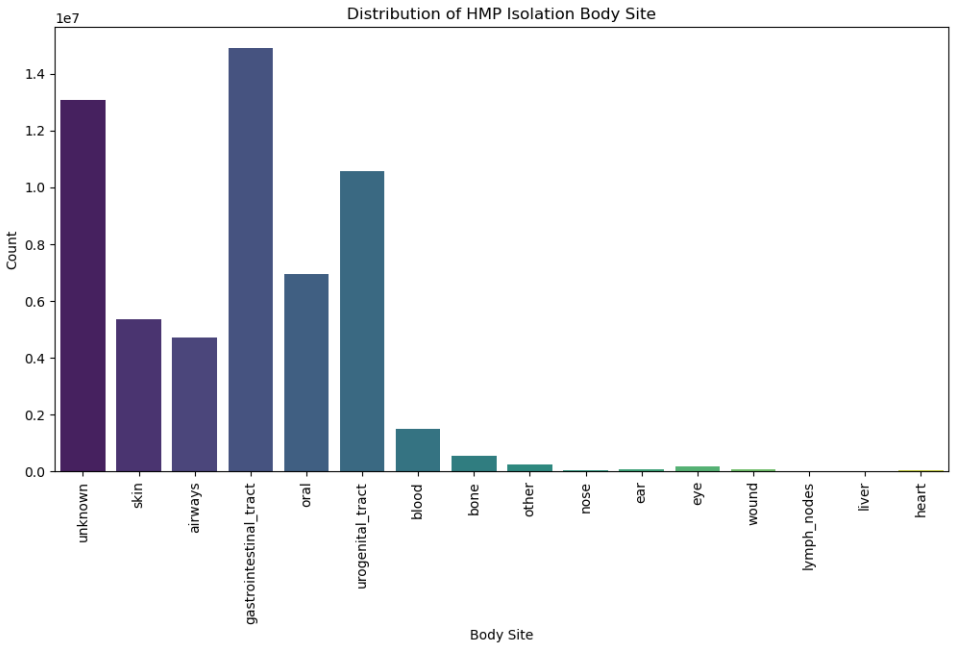


(a) *Original Data*

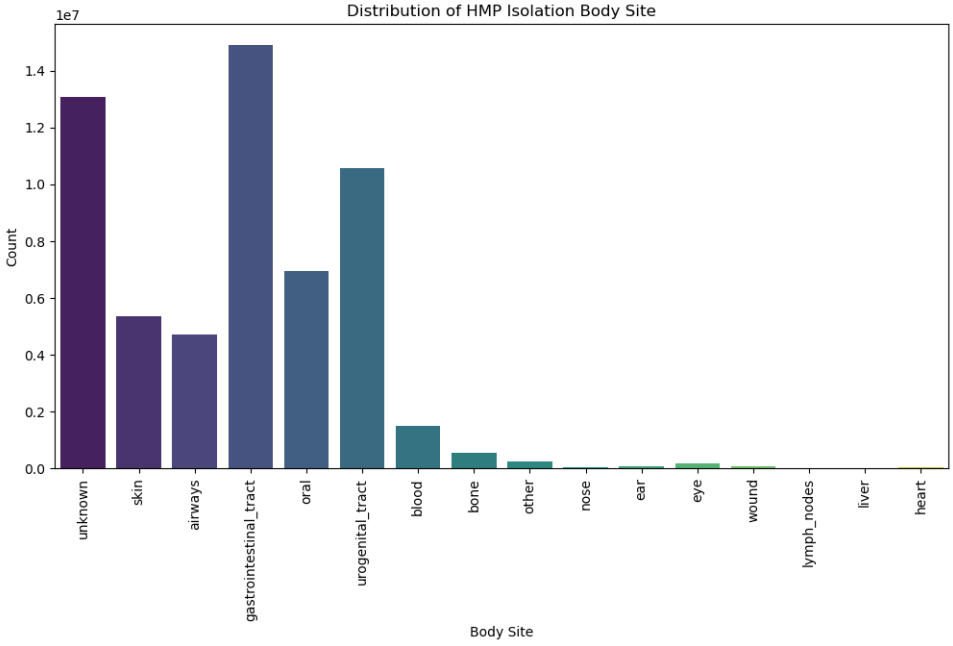


(b) *Augmented Data*

Figure 40: *HMP Body Site Distribution*

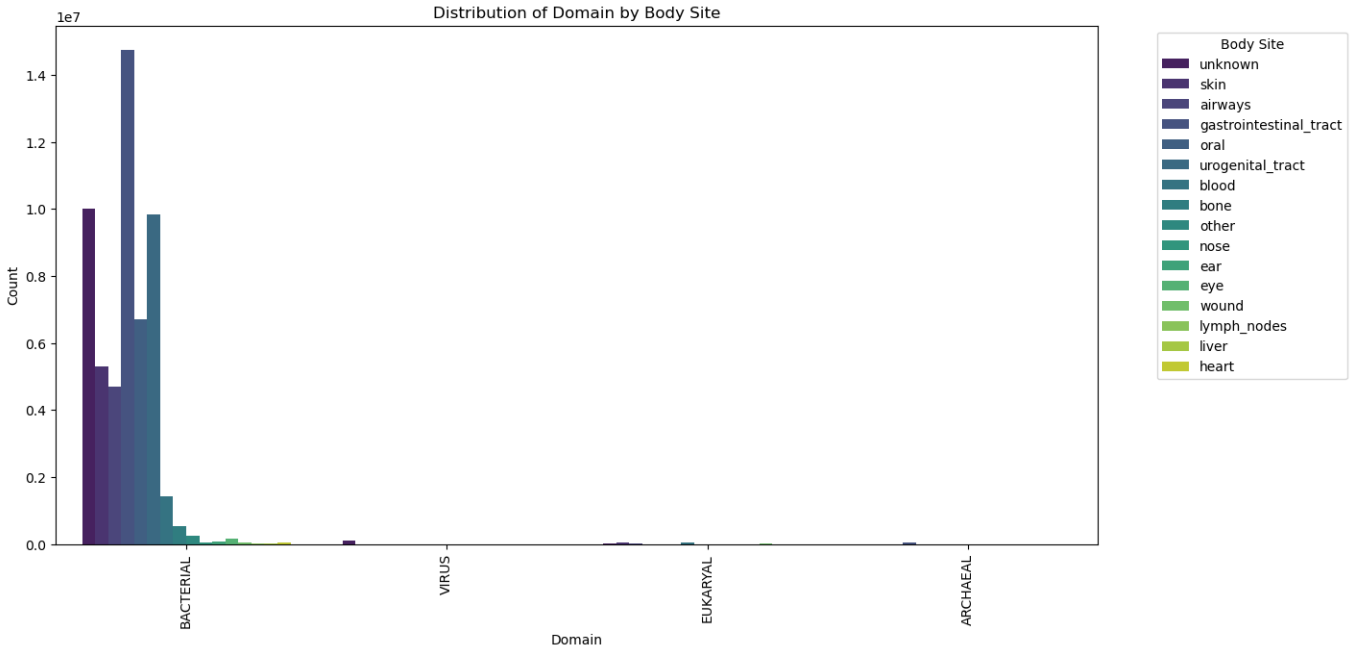


(a) *Original Data*

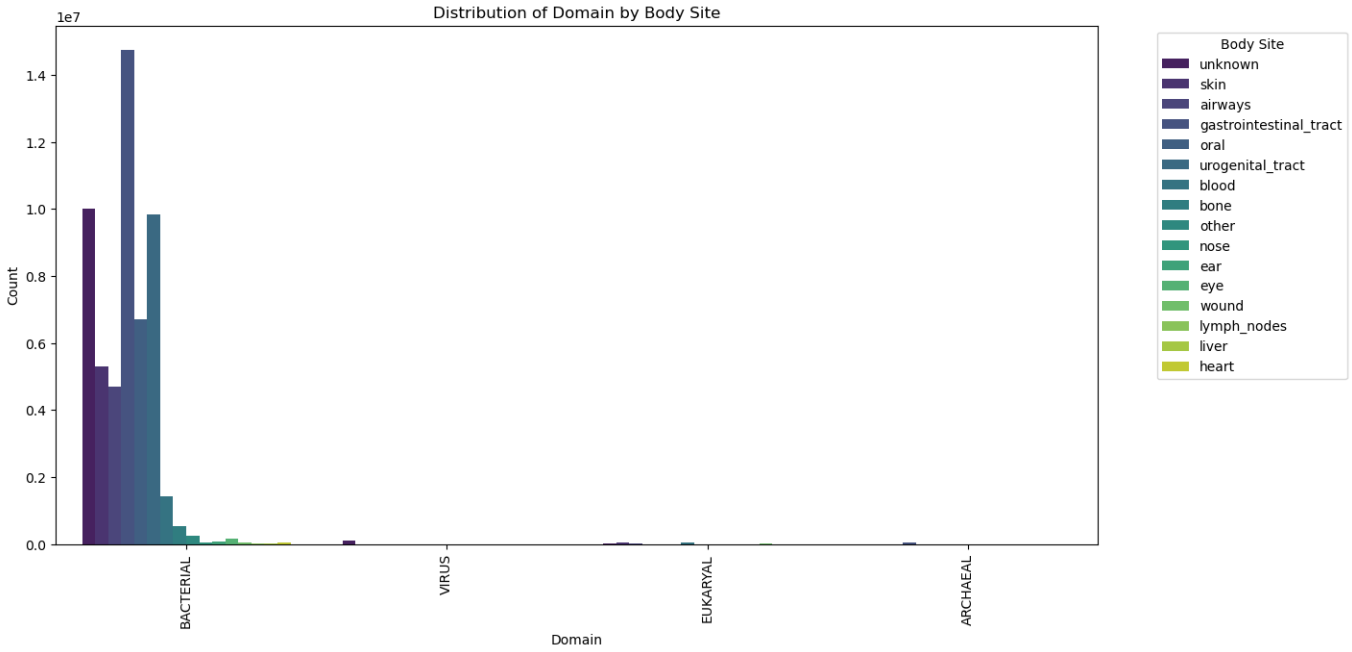


(b) *Augmented Data*

Figure 41: *Domain Distribution By Body Site*

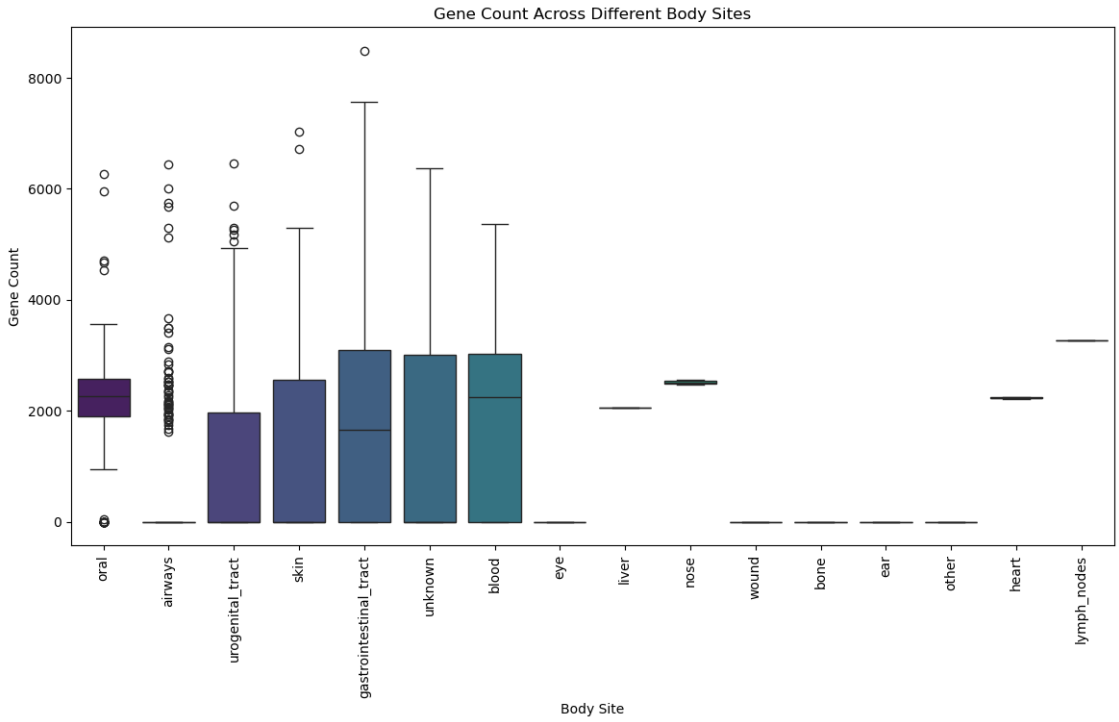


(a) *Original Data*

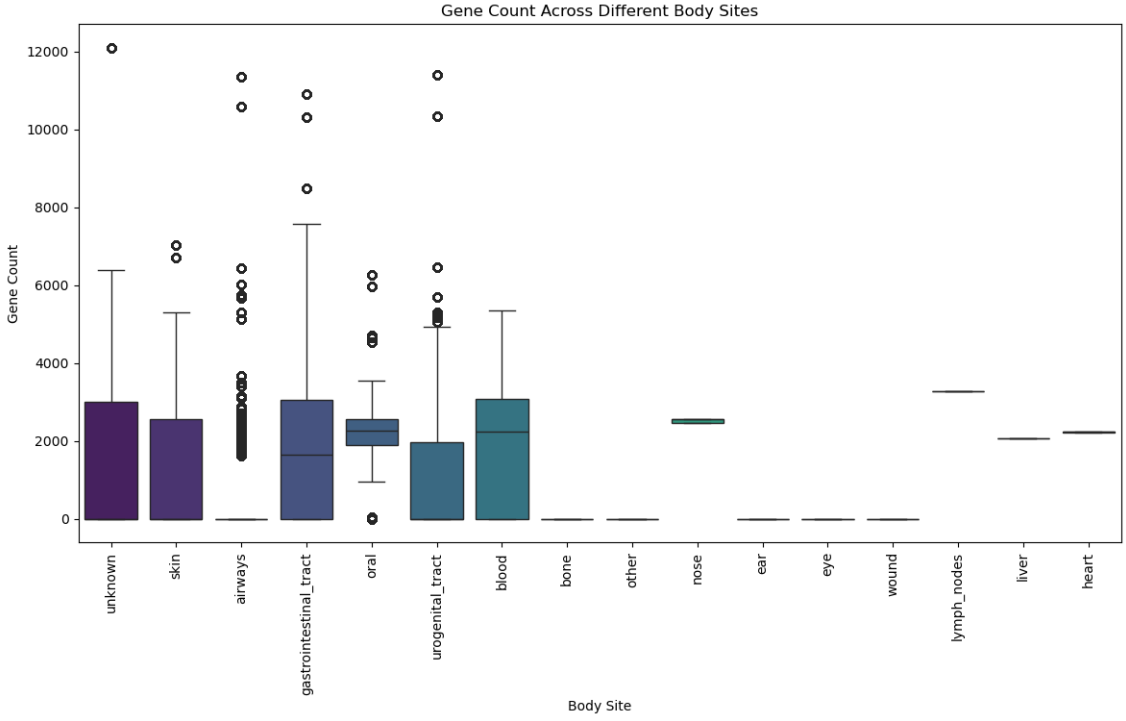


(b) *Augmented Data*

Figure 42: *Gene Count by Body Site*



(a) *Original Data*

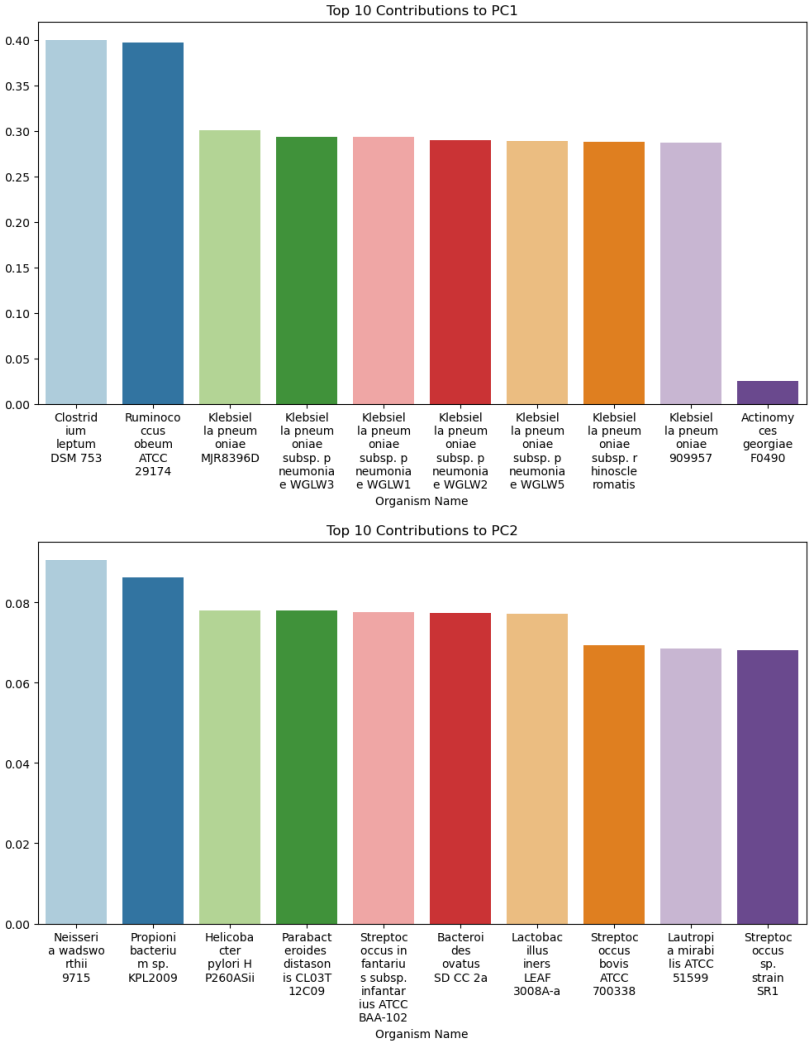


(b) *Augmented Data*

The gene count distribution across different body sites in both the original and augmented datasets highlights key variations. The augmented data reflects a broader distribution range, particularly in gastrointestinal and oral sites, while preserving the overall patterns observed in the original data.

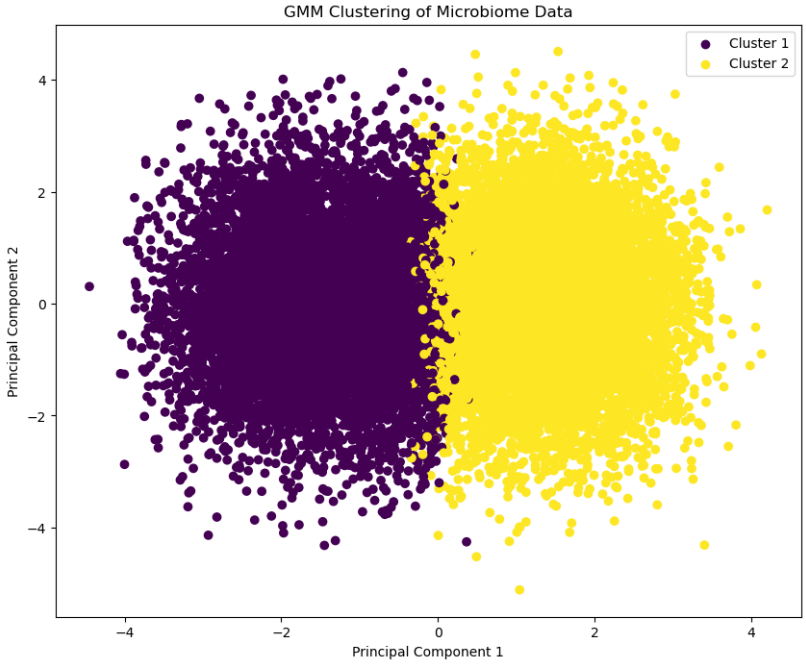
M Microbiome Clustering Figures

Figure 43: *Inspection of the Principal Components*



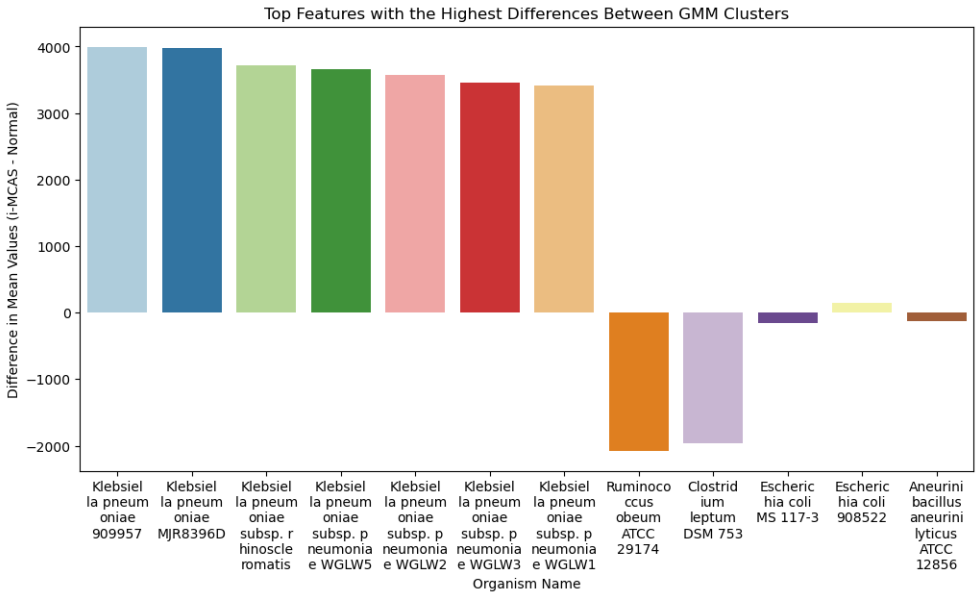
Source: Lea Brody-Heine, *Microbiome Models.ipynb*

Figure 44: *Gaussian Mixture Model Clusters*



Source: Lea Brody-Heine

Figure 45: *Gaussian Mixture Model Cluster Analysis*



Source: Lea Brody-Heine