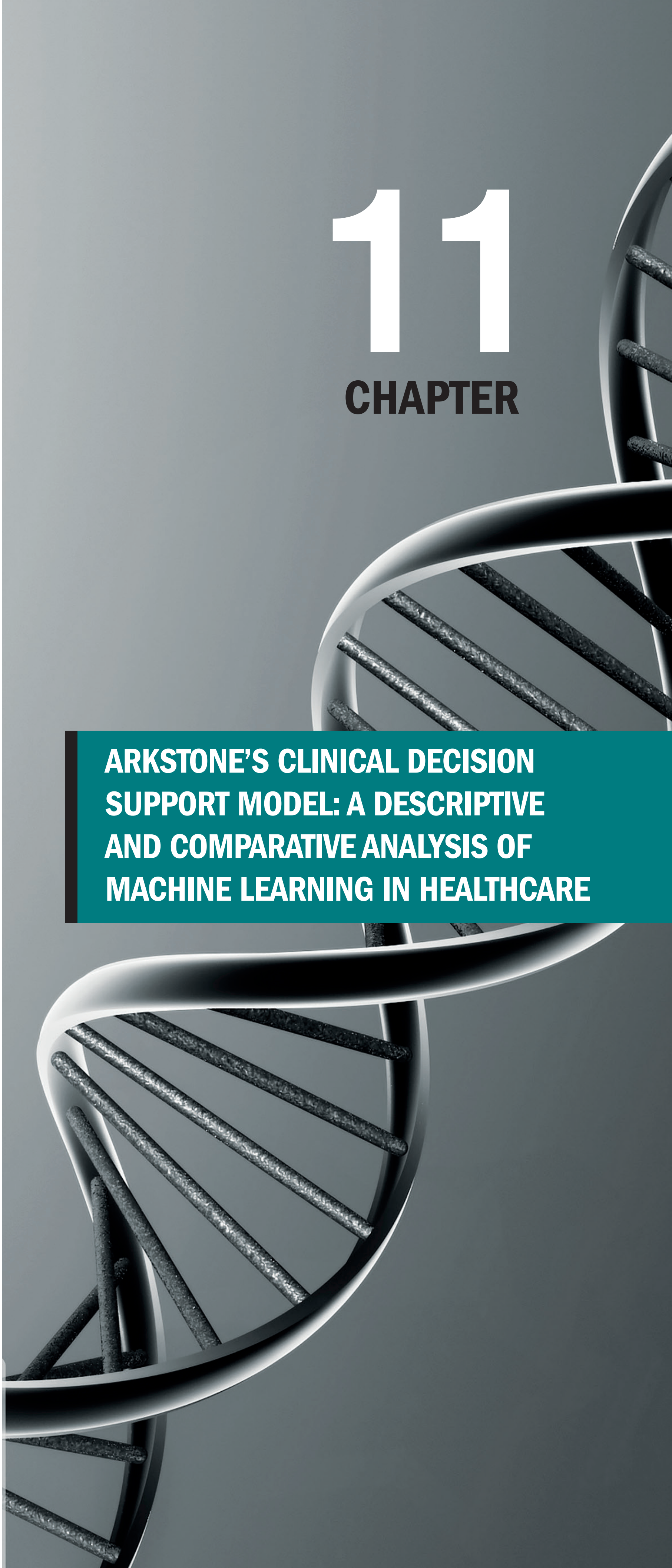

11

CHAPTER

ARKSTONE’S CLINICAL DECISION SUPPORT MODEL: A DESCRIPTIVE AND COMPARATIVE ANALYSIS OF MACHINE LEARNING IN HEALTHCARE

CHAPTER 11 ARKSTONE'S CLINICAL DECISION SUPPORT MODEL: A DESCRIPTIVE AND COMPARATIVE ANALYSIS OF MACHINE LEARNING IN HEALTHCARE

Machine Learning (ML) has significantly impacted various industries, including healthcare, by enabling systems to learn from data and enhance decision-making. However, deploying ML models in clinical settings presents challenges such as ensuring model reliability, avoiding overfitting, managing and receiving quality data, as well as maintaining data integrity and security. This paper provides a detailed analysis of Arkstone's clinical decision support system (CDSS) powered by machine learning. It explores the innovative validation techniques, training methodologies, and data management strategies employed by Arkstone to overcome common challenges in ML model deployment, particularly in providing clinical decision support to healthcare providers. Arkstone's approach integrates a human-in-the-loop (HITL) process, which ensures that every data input is meticulously reviewed by experts before being used for clinical decision-making. The paper also compares Arkstone's methodologies with traditional model validation frameworks, highlighting its contributions to improving accuracy, adaptability, and real-time performance in clinical environments. By focusing on data integrity, scalability, and model validation, Arkstone's system exemplifies a robust, secure, and effective solution for clinical decision-making. The findings aim to contribute to the broader discourse on achieving reliable and ethical AI systems in healthcare.

INTRODUCTION

ML and its Broad Applications

ML, a subset of artificial intelligence (AI), has profoundly transformed numerous industries by enabling systems to learn from data, recognize patterns, and make autonomous decisions with minimal human intervention **(Mitchell, 1997)**. This technological breakthrough has diverse applications, enhancing efficiency and accuracy in tasks traditionally reliant on manual processes. Industries such as healthcare, finance, autonomous transportation, and natural language processing have significantly benefited from ML driven innovations.

In healthcare, ML models are utilized to create predictive systems that aid in diagnosing diseases, tailoring personalized treatment plans, and managing patient care more effectively **(Esteva *et al.*, 2017)**. For example, predictive models leveraging electronic health records (EHRs) can forecast disease progression, thereby improving clinical outcomes **(Shickel *et al.*, 2018)**. In the financial sector, ML algorithms play a pivotal role in detecting fraudulent transactions, forecasting market trends, and optimizing portfolio management strategies **(Feng *et al.*, 2019)**. Autonomous vehicles rely heavily on ML for navigating complex environments, ensuring passenger safety, and enhancing transportation efficiency **(Bojarski *et al.*, 2016)**. Similarly, natural language processing (NLP) applications—such as machine translation, sentiment analysis, and conversational agents—rely on advanced ML models to interpret and generate human language **(Brown *et al.*, 2020)**.

Challenges in ML Model Deployment

Despite its transformative potential, deploying ML models in real-world applications presents substantial challenges. Ensuring reliability, generalizability, and robust performance on unseen data remains a critical requirement. Rigorous validation processes are essential to avoid overfitting—where a model captures noise or spurious patterns in the training data—and underfitting, where a model oversimplifies data patterns due to insufficient complexity or poor data representation **(Hastie *et al.*, 2009)**.

Traditional validation techniques, including K-Fold Cross-Validation, Leave-One-Out Cross-Validation, and Stratified K-Fold Cross-Validation, are widely employed to evaluate model performance **(Kohavi, 1995)**. These methods offer structured approaches for hyperparameter tuning and model selection, thereby mitigating issues such as overfitting and bias. However, they also encounter limitations, particularly concerning data quality, computational efficiency, and comprehensive model assessment. In addition, many of these validation techniques were designed to evaluate older models, and applying these approaches to newer systems falls short of comprehensively evaluating the capabilities and accuracy of the model.

One of the greatest obstacles to effective ML is the availability and quality of diverse, representative datasets **(Shickel *et al.*, 2018)**. Poor data quality introduces biases and errors, compromising model accuracy and generalizability. For instance, healthcare models trained on skewed datasets may fail to generalize to broader populations, leading to erroneous predictions **(Obermeyer *et al.*, 2019)**. Additionally, systems relying exclusively on static, pre-trained data often struggle to adapt to evolving data patterns, resulting in model drift and reduced performance over time. Retraining models to account for new information presents further challenges in terms of cost, data availability, and computational demands **(Gama *et al.*, 2014)**.

Security and privacy concerns also present formidable barriers. Sensitive data, especially in healthcare and finance, must comply with stringent privacy regulations, including GDPR and HIPAA **(Voigt & Von dem Bussche, 2017)**. Balancing the need for extensive training data with privacy preservation remains a key concern for ML practitioners.

This paper aims to contribute to the broader discourse on ML model validation, emphasizing practical solutions for achieving robust, secure, and ethical AI systems.

Nomenclature Explained at Length:

Clinical Decision Support Software

Clinical Decision Support Systems (CDSS) powered by ML have gained prominence for their potential to improve healthcare outcomes. These systems assist clinicians by analyzing extensive medical data to provide evidence-based recommendations. A critical element of CDSS is the rigorous validation of ML models to ensure accuracy, reliability, and applicability in clinical environments. Accurate model validation is particularly important in high-stakes medical diagnostics, where errors can lead to severe consequences **(Cabitza, Rasoini, & Gensini, 2017)**. Additionally, data quality, model bias, and interpretability are persistent challenges **(Mehrabi *et al.*, 2021; Rudin, 2019)**.

Non-Device CDSS vs Device CDSS

The distinction between Non-Device Clinical Decision Support Software (CDSS) and Device CDSS is based on regulatory criteria set by the FDA (fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device). To be classified as a Non-Device CDSS, the software must meet all four of the following conditions:

1. It does not acquire, process, or analyze medical images, signals, or patterns.
   - Example: Software that only displays medical information normally communicated between healthcare professionals (HCPs).
   - Device CDSS, on the other hand, involves processing medical signals, such as continuous glucose monitoring (CGM) data or ECG waveforms.
2. It displays, analyzes, or prints medical information that is already well understood in clinical decision-making.
   - Example (Non-Device): Displaying a single test result that is already clinically meaningful.
   - Device CDSS may analyze complex signals or patterns requiring interpretation.
3. It provides recommendations rather than specific outputs or directives.
   - Example (Non-Device): A list of possible treatment options based on clinical guidelines.
   - Device CDSS may generate time-critical outputs or directives, such as automated diagnosis or required actions.
4. It provides the basis for recommendations so that an HCP does not rely primarily on the software's output.
   - Example (Non-Device): The software provides transparent reasoning, including how it derives conclusions.
   - Device CDSS might not disclose the basis of recommendations, making the HCP rely more heavily on the software's output.

Table 1.

Criteria	Non-Device CDSS	Device
Medical Data Processing	Does not acquire, process, or analyze signals, images, or patterns	Processes or analyzes medical signals, images, or patterns
Type of Information Handled	Displays and prints clinically understood information	Involves continuous signals, imaging, and diagnostic patterns
Output Type	Provides recommendations or options for decision-making	May generate specific, time-critical outputs or directives
Role of HCP in Decision-Making	HCP does not rely solely on the software; reasoning is transparent	HCP may rely primarily on software's output without full transparency
Examples	Lists of treatment options, clinical guidelines, reference information	MRI interpretation, continuous glucose monitoring, ECG analysis

Predictive ML vs Non-Predictive ML

ML models can be categorized primarily into predictive and non-predictive models (www.healthit.gov/sites/default/files/page/2023-04/NPRM_DSI_fact%20sheet-508.pdf). While both types aim to extract valuable insights from data, they differ significantly in their objectives, methodologies, and applications. Predictive models in ML are primarily designed to forecast future outcomes or classify data based on patterns learned from historical data. These models are often trained using supervised learning, a method where the model learns from labeled data to make predictions about unseen data. Common predictive models include linear regression, decision trees, support vector machines, and neural networks **(Bishop, 2006)**. The performance of these models is typically assessed using metrics such as accuracy, precision, recall, F1-score, and mean squared error, which evaluate how well the model predicts or classifies new data **(Jordan & Mitchell, 2015)**.

On the other hand, non-predictive models in ML are designed to uncover hidden patterns or structures within data without the need for labeled outcomes. These models are typically employed in unsupervised learning, where the system attempts to identify inherent relationships or groupings in the data. Non-predictive models are particularly useful for tasks like clustering, dimensionality reduction, and association rule mining. Techniques such as K-means clustering, principal component analysis (PCA), and the Apriori algorithm are commonly used in these models **(Hastie, Tibshirani, & Friedman, 2009)**. Unlike predictive models, which focus on predicting outcomes, non-predictive models seek to explore and understand the data itself. The evaluation of non-predictive models is more focused on how well they uncover meaningful data structures, with common metrics including silhouette scores for clustering and explained variance for dimensionality reduction **(Hastie *et al.*, 2009)**.

The data used in these two types of models further distinguishes them. Predictive models depend on labeled data, where each input feature corresponds to a known outcome. This labeled data allows the model to learn from historical examples and apply that knowledge to predict future instances. For example, a predictive model trained on a dataset of patients with labeled outcomes (e.g., whether they developed a certain disease) can predict whether a new patient will develop the disease based on their health information. In contrast, non-predictive models generally operate on unlabeled data, seeking to identify patterns without prior knowledge of the outcomes. For instance, a non-predictive model might group customers into segments based on their purchase behavior, without knowing in advance what those segments might represent.

Table 2.

	Predictive ML	Non-Predictive ML CDSS
Objective	Forecasts future outcomes or classifies data based on historical patterns.	Focuses on analyzing current data without making future predictions, emphasizing accurate recommendations.
Learning Type	Often employs supervised learning, with models trained on labeled data to predict unseen outcomes.	Avoids predictions; uses structured data analysis and predefined rules for recommendations.
Common Techniques	Regression, decision trees, support vector machines, neural networks.	Rule-based systems, data validation frameworks, expert-guided outputs.
Evaluation Metrics	Accuracy, precision, recall, F1-score, mean squared error.	Quality of recommendations, consistency with expert guidelines, and clinical accuracy.
Data Requirements	Requires labeled datasets where each input corresponds to an output.	Relies on validated inputs and expert-reviewed outputs rather than learning from unlabeled data.
Applications	Disease prediction, fraud detection, financial forecasting.	Clinical decision support systems, diagnostic assistance, and evidence-based recommendations.
Arkstone Approach	Does not use predictive modeling; focuses exclusively on providing recommendations validated by humans.	Relies on structured data validation and expert oversight to ensure accuracy and reliability.

Supervised ML vs Unsupervised ML

ML models can also be broadly classified into supervised and unsupervised learning. These two paradigms differ in terms of their objectives, data usage, and techniques. Supervised learning is primarily focused on making predictions or classifications using labeled data, while unsupervised learning is concerned with identifying patterns or structures in unlabeled data. Supervised learning is a ML paradigm where the model is trained on a labeled dataset, meaning each input in the dataset is paired with the correct output. The primary goal of supervised learning is to make predictions or classifications based on historical data. By learning from these labeled examples, the model can generalize to new, unseen data and provide accurate predictions. Common techniques used in supervised learning include regression models, decision trees, support vector machines, and neural networks **(Bishop, 2006)**. Unsupervised learning, in contrast, works with unlabeled data. The objective of unsupervised learning is to uncover hidden structures, patterns, or relationships within the data without prior knowledge of the output. Unsupervised learning algorithms attempt to group similar data points together or reduce the dimensionality of the data. One of the key differences between supervised and unsupervised learning lies in the type of data used. Supervised learning relies on labeled data, where each input is associated with a known output. This type of data is often collected in scenarios where the desired outcome is already known or can be easily categorized. In contrast, unsupervised learning operates on unlabeled data.

Another distinguishing factor is the way the models are evaluated. In supervised learning, evaluation is relatively straightforward, as the model's predictions can be compared directly to the known labels in the dataset. Common performance metrics include accuracy, precision, recall, and F1-score for classification tasks, or mean squared error (MSE) for regression tasks **(Jordan & Mitchell, 2015)**. The evaluation process in supervised learning involves assessing how well the model generalizes to unseen data, ensuring that it can make accurate predictions in real-world scenarios. In unsupervised learning, evaluating model performance is more complex because there is no ground truth to compare the results against. Instead, evaluation typically focuses on how well the model has identified meaningful patterns or structures in the data. For clustering tasks, metrics like silhouette score or adjusted Rand index are used to measure the cohesion and separation of clusters, while methods like explained variance are used for dimensionality reduction **(Hastie *et al.*, 2009)**.

Table 3.

	Supervised ML	Unsupervised ML
Objective	Makes predictions or classifications based on labeled training data.	Identifies patterns, relationships, or structures in unlabeled data.
Data Requirements	Requires labeled data with known input-output pairs.	Works with unlabeled data without predefined categories or labels.
Common Techniques	Regression models, decision trees, support vector machines, neural networks.	Clustering (e.g., K-means), dimensionality reduction (e.g., PCA), hierarchical clustering.
Evaluation Metrics	Accuracy, precision, recall, F1-score, mean squared error (regression).	Silhouette score (clustering), explained variance (dimensionality reduction).
Applications	Predictive analytics, fraud detection, diagnostic tools.	Market segmentation, exploratory data analysis, anomaly detection.
Arkstone Approach	Utilizes supervised techniques to validate all inputs through human oversight.	Avoids traditional unsupervised methods; focuses instead on structured validation to ensure accuracy.

OVERVIEW OF OTHER CLINICAL DECISION SUPPORT TOOLS

IBM Watson for Oncology IBM Watson for Oncology, launched in 2014 through collaboration with Memorial Sloan Kettering Cancer Center (MSKCC), aimed to enhance oncology care by providing evidence-based treatment recommendations derived from vast data, including patient records, medical literature, and clinical guidelines **(Ross & Swetlitz, 2017)**. The system sought to deliver personalized, research-supported treatment plans, improve efficiency by reducing time spent reviewing extensive information, and continuously adapt by integrating new data **(Howard & Shapiro, 2018; Strickland, 2019)**. Watson for Oncology employed natural language processing (NLP) and ML to assist in cancer treatment, with its recommendations validated through retrospective comparisons to decisions from multidisciplinary tumor boards (MDT) **(Zhou *et al.*, 2018)**.

Despite its promising objectives, Watson for Oncology ultimately failed to achieve its goals and was discontinued in 2020. One of the primary reasons for its failure was the accuracy and reliability of its recommendations. The system was often criticized for providing inaccurate and sometimes unsafe treatment options that were not clinically validated or aligned with standard practices **(Herper, 2018)**. This issue was compounded by data limitations, as Watson for Oncology relied heavily on information from MSKCC, which might not have been representative of the broader patient population. Consequently, the recommendations lacked generalizability **(Miliard, 2020)**.

The inherent complexity and variability of cancer treatment posed significant challenges for the AI, which struggled to adapt to the nuances of different cases **(Winkler, 2018)**. Moreover, many oncologists were skeptical about relying on an AI system for treatment decisions, preferring their clinical judgment and experience **(Farr, 2018)**. Integration challenges further hindered the adoption of Watson for Oncology, as it proved difficult to integrate the system into existing healthcare workflows and electronic health record systems **(Klein, 2019)**.

IBM's marketing of Watson for Oncology often oversold its capabilities, leading to high expectations that the system couldn't meet **(Strickland, 2019)**. These issues, combined with the complex nature of cancer treatment and the skepticism from medical professionals, led to the eventual discontinuation of Watson for Oncology. IBM refocused its Watson Health efforts and sold parts of the business to other companies, marking the end of an ambitious but ultimately unsuccessful venture **(Miliard, 2020)**.

DeepMind Health's Streams:

DeepMind Health Stream utilizes deep learning to detect acute kidney injury (AKI) early and its validation process includes clinical trials and real-world deployment, emphasizing early warning capabilities **(Niel *et al.*, 2018)**.

DeepMind Health's Streams was an ambitious project developed by DeepMind, a subsidiary of Alphabet, to revolutionize healthcare by leveraging advanced ML and artificial intelligence. Launched in 2016, Streams aimed to provide clinicians with real-time alerts and insights to improve patient outcomes, specifically focusing on early detection of acute kidney injury (AKI) and other conditions. The primary objective was to create a mobile app that would streamline the flow of critical patient information, enabling healthcare professionals to respond swiftly and effectively to signs of deterioration **(Hern, 2017)**.

The Streams app sought to address a critical gap in the timely identification and management of serious health conditions. By processing data from electronic health records (EHRs), the app was designed to send instant alerts to clinicians, highlighting patients at risk of AKI, a condition often missed in its early stages. The overarching goal was to reduce the incidence of avoidable complications and hospital admissions by ensuring that medical staff had access to actionable information at their fingertips **(Powles & Hodson, 2017)**.

Despite the noble objectives and initial promise, Streams faced significant challenges that ultimately led to its failure. One of the primary issues was related to data privacy and governance. The initial partnership with the Royal Free London NHS Foundation Trust involved sharing patient data without explicit patient consent, which led to concerns and backlash regarding the handling of sensitive medical information. This controversy sparked debates about data protection and the ethical implications of using patient data for AI-driven healthcare solutions **(Kelion, 2017)**.

Another significant challenge was the integration of the Streams app into the complex and often fragmented healthcare systems. Ensuring seamless interoperability with existing EHR systems proved to be a daunting task, hindering the widespread adoption of the technology. Additionally, while the app showed potential in improving patient care for specific conditions like AKI, expanding its functionality to cover a broader range of medical issues proved more complex than anticipated **(Creswell, 2019)**.

Furthermore, the transition of DeepMind Health's operations to Google Health in 2018 added to the project's challenges. This shift raised additional concerns about data privacy and the commercialization of patient information, as stakeholders worried about how Google might use the data **(Wagner, 2018)**.

Ultimately, these issues of data privacy, integration challenges, and the complexities of expanding the app's functionalities contributed to the discontinuation of Streams. While the project demonstrated the potential of AI in transforming healthcare, it also highlighted the significant hurdles that must be overcome to integrate advanced technologies into medical practice ethically and effectively **(Hern, 2019)**.

Mayo Clinic's Predictive Analytics:

Mayos Clinics Predictive Analytics employed ML algorithms to predict patient deterioration in the ICU. Its validation included retrospective cohort studies and comparison with standard ICU protocols **(Komorowski *et al.*, 2018)**.

Mayo Clinic's Predictive Analytics initiative was an effort to leverage advanced data analytics and ML techniques to improve patient care and operational efficiency. Launched with high expectations, the project aimed to harness the vast amounts of patient data available at Mayo Clinic to predict patient outcomes, optimize treatment plans, and streamline hospital operations. The primary objective was to use predictive models to anticipate clinical events such as patient deterioration, readmissions, and complications, thereby enabling preemptive interventions and enhancing overall patient outcomes **(Kharbanda, 2019)**.

One of the main goals of the Predictive Analytics initiative was to integrate predictive insights into the daily workflows of healthcare providers. By doing so, the system hoped to provide real-time decision support that could alert clinicians to potential issues before they became critical. This proactive approach was expected to reduce the incidence of adverse events, improve patient safety, and decrease healthcare costs. Additionally, the initiative aimed to optimize resource allocation within the hospital, such as better management of bed occupancy and staffing levels, by predicting patient flow and demand **(Topol, 2019)**.

Despite its ambitious objectives, the Predictive Analytics project faced several challenges that ultimately led to its failure. One significant issue was the complexity and variability of clinical data. The models often struggled with the accuracy and reliability of predictions due to the diverse and nuanced nature of patient conditions. Moreover, integrating these predictive tools into existing clinical workflows proved to be more challenging than anticipated, as it required significant changes in how healthcare providers operated and made decisions **(McKinney, 2020)**.

Data privacy and governance also posed substantial challenges. Ensuring the secure handling of sensitive patient information while maintaining the accuracy and utility of predictive models was a complex task. There were concerns about the ethical implications of using patient data for predictive purposes, which led to resistance from both clinicians and patients **(Jiang *et al.*, 2017)**.

Additionally, there was skepticism among healthcare providers regarding the reliability of the predictive models. Many clinicians were hesitant to rely on algorithmic recommendations over their clinical judgment and experience. This skepticism hindered the adoption and effective use of the predictive tools, limiting their impact on patient care **(Topol, 2019)**. Ultimately, the combination of data complexity, integration challenges, privacy concerns, and clinician skepticism led to the discontinuation of Mayo Clinic's Predictive Analytics initiative. While the project demonstrated the potential benefits of predictive analytics in healthcare, it also highlighted the significant hurdles that must be overcome to implement such technologies successfully **(Kharbanda, 2019)**.

PEDIATRIC ALERT SYSTEM

In another study of clinical decision support software in pediatrics, multiple validation methods, such as cross-validation, were used to estimate the machine's generalization performance by testing it on unseen data **(Ramgopal *et al.*, 2022)**. Additionally, sensitivity and specificity calculations were employed to assess the effectiveness of these validation methods **(Ramgopal *et al.*, 2022)**.

The researchers aimed to validate specific components of their ML system, including the algorithm used to develop the AI-CDS system. They assessed whether the inputted information could accurately and effectively identify patients at risk for specific diseases and evaluated the model's ability to predict disease outcomes with few false alerts compared to traditional CDS models in the pediatric population **(Ramgopal *et al.*, 2022)**. External validation was used to determine the new model's reliability and its capability to generalize predictions beyond specific datasets. Through this validation, researchers evaluated whether the model maintained its predictive performance, accuracy, and safety across populations **(Ramgopal *et al.*, 2022)**. They also inspected inaccurate predictions for systematic errors and evaluated the model's accuracy for vulnerable subgroups to achieve the goal of integrating the AI-CDS into the healthcare system **(Ramgopal *et al.*, 2022)**.

Although the validation methods used in the studies were indeed valuable, some pitfalls highlighted included the lack of evidence-based guidelines and variations in care. CDS is most effective when disease detection can promote evidence-based care. However, the authors concluded that the lack of large datasets places limitations on the capabilities of this particular CDS model **(Ramgopal *et al.*, 2022)**.

MODEL VALIDATION TECHNIQUES

Model validation, the process of evaluating a ML model's performance on independent data, is essential for addressing these concerns. Validation helps ensure that a model not only performs well on training data but also maintains its accuracy and robustness when exposed to new data **(Yao, Rosasco, & Caponnetto, 2007)**. Proper validation techniques can reveal overfitting, where a model learns noise and patterns specific to the training data, leading to poor generalization. Furthermore, validation is crucial for hyperparameter tuning and model selection, providing a framework to systematically compare different models and configurations to identify the best-performing one **(Bergstra & Bengio, 2012)**.

Model validation is a crucial step in the development and deployment of ML models, particularly in healthcare, where the reliability and accuracy of predictions can significantly impact patient outcomes. Several techniques are commonly used to validate these models, each with its own advantages and applications.

Cross-validation is a widely used technique where data is split into training and test sets multiple times to ensure that model performance is robust across different subsets. This method helps in assessing the stability and generalizability of the model. For instance, cross-validation is commonly used in studies validating ML models for sepsis prediction and diabetic retinopathy screening **(Rajkomar *et al.*, 2018)**. K-Fold Cross-Validation involves dividing the dataset into K equal-sized folds. The model is trained K times, each time using K-1 folds for training and the remaining fold for validation. The performance metrics are then averaged across all K iterations to obtain an overall estimate of the model's performance **(James *et al.*, 2013)**. Stratified K-Fold Cross-Validation is a variation of K-Fold Cross-Validation that ensures each fold preserves the proportion of classes in the original dataset. This is particularly useful for imbalanced datasets where certain classes are underrepresented. By maintaining class balance in each fold, stratified K-Fold Cross-Validation provides more reliable performance estimates, especially for classification tasks **(Raschka & Mirjalili, 2019)**. Leave-One-Out Cross-Validation (LOOCV) is a special case of K-Fold Cross-Validation where K equals the number of samples in the dataset. In LOOCV, the model is trained K times, each time using all but one sample for training and the remaining sample for validation. LOOCV provides a high-variance estimate of the model's performance but can be computationally expensive, especially for large datasets **(Hastie, Tibshirani, & Friedman, 2009)**.

Repeated Random Train-Test Splits involve randomly partitioning the dataset into training and testing sets multiple times. Unlike K-Fold Cross-Validation, the data splitting is not systematic, and each iteration may result in different training and testing sets. This approach is useful when computational resources are limited or when the dataset is too large to be efficiently processed using cross-validation techniques **(Raschka & Mirjalili, 2019)**.

Prospective validation involves testing the model in a real-time clinical environment to assess its performance in real-world settings. This method provides insights into how the model will function in practice, beyond controlled experimental conditions. A notable example of prospective validation is Google Health's AI for detecting diabetic retinopathy, which showed high sensitivity and specificity in clinical trials **(Abràmoff *et al.*, 2018)**. This approach ensures that the model performs well when integrated into actual clinical workflows.

External validation uses data from institutions not involved in the model's development to test its generalizability. This technique is essential for confirming that a model can be applied broadly across different populations and healthcare settings. Mayo Clinic's predictive models, for example, often undergo external validation to ensure their applicability and reliability in diverse healthcare environments **(Lundberg *et al.*, 2018)**.

Retrospective validation, on the other hand, involves analyzing historical data to evaluate model performance post hoc. This approach compares machine learning-generated recommendations with past clinician decisions to assess accuracy and effectiveness. IBM Watson for Oncology and other clinical decision support systems frequently use retrospective validation to benchmark their recommendations against historical clinical outcomes **(Esteva *et al.*, 2017)**.

Each of these validation techniques plays a critical role in ensuring that ML models are accurate, reliable, and generalizable across different clinical scenarios. By employing a combination of these methods, researchers and healthcare providers can better understand the strengths and limitations of their models, leading to more effective and trustworthy healthcare applications.

Understanding the Arkstone ML Model:

Arkstone is a sophisticated clinical decision support system designed to assist healthcare professionals in making informed decisions regarding the treatment of infectious diseases. Arkstone meets critera of a non-device CDSS as outlined by the FDA(https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software). This is because Arkstone does not anaylze, acquire or process medical images, signals or patterns, functions to display anylze medical information normally communicated between healthcare provides (lab results and demographics), and functions to provide recommendations to healthcare providers rather than a specific output or directive. In addition, healthcare providers have access to the basis of the recommendations and the primary sources in which the recommendations were derived from, as well as insight into the inputs used to generate the recommendations including descriptions of the ML processes and validation techniques. The platform leverages advanced supervised ML techniques to offer expert guidance tailored to the unique needs of each patient. However, unlike many other ML systems, Arkstone operates as a non-predictive model, requiring all inputs and outputs to undergo rigorous validation processes. Every step, from the entry of data to the outputed recommendation put forth to clinicians, is subject to human approval, ensuring that all data is meticulously trained and aligned with already established outputs. This structured approach ensures that Arkstone never generates predictive outputs independent of human approval and instead focuses on providing guidance based on well-validated, known outcomes.

DATA INPUTS AND INTEGRITY

Arkstone receives data through three primary channels: API integration with Laboratory Information Systems (LIS) or Laboratory Information Management Systems (LIMS), HL7

Table 4.

Validation Technique	Description	Advantages	Limitations	Example Applications
Cross-Validation	Splits data into training and test sets multiple times to assess model robustness across subsets.	Robust assessment of stability and generalizability.	Can be computationally intensive for large datasets.	Commonly used for sepsis prediction and diabetic retinopathy screening.
K-Fold Cross-Validation	Divides the dataset into K equal-sized folds, training on K-1 folds and validating on the remaining fold.	Provides a comprehensive performance estimate by averaging results across folds.	May be less effective for very small datasets.	General-purpose ML model validation.
Stratified K-Fold Cross-Validation	Ensures each fold preserves the class proportions in the original dataset.	Suitable for imbalanced datasets, providing more reliable estimates for classification tasks.	Limited to classification tasks with categorical data.	Commonly used for classification models with imbalanced datasets.
Leave-One-Out Cross-Validation (LOOCV)	A special case of K-Fold where K equals the number of samples; trains on all but one sample at a time.	Maximizes use of the dataset and gives high variance estimates.	Extremely computationally expensive for large datasets.	Extremely computationally expensive for large datasets.
Repeated Random Train-Test Splits	Randomly partitions data into training and test sets multiple times without systematic splits.	Useful when computational resources are limited or for large datasets.	Results may vary significantly due to the randomness of splits.	Useful for testing large datasets efficiently.
Prospective Validation	Tests the model in real-world clinical environments to assess real-time performance.	Provides insights into real-world usability and integration into workflows.	Time-intensive and challenging to control external variables.	Google Health's diabetic retinopathy detection model.
External Validation	Uses data from institutions not involved in the model's development to test generalizability.	Confirms the model's applicability across different populations and settings.	Requires access to external data, which may not always be available.	Mayo Clinic's predictive models for healthcare settings.
Retrospective Validation	Analyzes historical data to compare model predictions with past clinical decisions.	Allows benchmarking against established practices; useful for post hoc analysis.	Dependent on the quality of historical data and may not account for changes in clinical practices.	IBM Watson for Oncology and other clinical decision support systems.

integration with these same systems, and manual data entry via the Arkstone portal. The data transferred to Arkstone includes detailed information from laboratory results, including the methods of detection (such as cultures and/or molecular techniques), as well as patient demographics. These demographics encompass a range of details such as patient allergies, age, gender, pregnancy status, and diagnosis codes.

Once the data reaches Arkstone, it is protected by a strict integrity framework: once data is transmitted, it cannot be altered and protected internally from any alterations. This ensures the system maintains the highest standards of data accuracy and reliability. Arkstone processes thousands of lab results daily, sourced from across the United States and internationally, creating a rich and diverse dataset. These inputs come from a wide array of medical specialties, ranging from small independent practices to large corporate healthcare systems. The patient demographics processed by Arkstone can span the entire lifecycle, from neonates to geriatrics, reflecting the system's adaptability to different age groups and medical conditions.

In addition, the diverse nature of the real-life data that Arkstone processes plays a critical role in preventing issues like overfitting or underfitting in the model. Overfitting occurs when a model becomes too tailored to the training data, capturing noise or irrelevant patterns that do not generalize well to new data. Underfitting, on the other hand, happens when a model is too simplistic and fails to capture the underlying patterns in the data. By incorporating data from a wide range of sources, specialties, and patient demographics, Arkstone ensures that its machine-learning model is exposed to a broad spectrum of cases and scenarios. This diversity prevents the model from becoming overly specialized to any single dataset, ensuring that it maintains its ability to make accurate, generalized decisions across varied clinical situations.

Moreover, Arkstone is built to adapt to the ever-evolving landscape of healthcare data. As new information becomes available—whether it pertains to rare or newly described microbes, emerging allergies, or cutting-edge diagnostic technologies—the system is designed to integrate this data seamlessly, without imposing limitations on the types of variables it can process. This flexibility allows Arkstone to stay up-to-date with the latest developments in medical research and clinical practice. It also allows the labs to tailor their offerings to their specific needs and the clinician's specific needs ather than creating lab offerings tailored to predetermined recommendations, bypassing any biases that would occur if only a finite amount of variables can be processed by Arkstone. For example, if a new pathogen is identified or if a novel diagnostic technique is introduced, Arkstone can incorporate this data into its decision-making process, ensuring that clinicians have access to the most current and relevant information. Therefore, there are no restrictions on the nature of patient demographics, the number or types of organisms detected, the detection of resistance genes or sensitivity patterns, or combinations of resistance genes and organisms identified. Additionally, there are no limitations on the number or combination of allergies detected, or on the type, size, combination, or quantity of variables that can be processed. This adaptability is crucial for maintaining the accuracy and effectiveness of the system as medical knowledge continues to evolve and expand, ensuring that Arkstone remains a valuable and reliable tool for clinicians in the face of rapidly changing healthcare environments.

Arkstone places a high priority on data security and privacy, adhering to the stringent standards set forth by healthcare regulatory bodies such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States. These regulations govern how patient information must be handled, stored, and transmitted to ensure that it remains confidential and protected from unauthorized access. Arkstone follows these standards rigorously, maintaining servers that meet industry benchmarks for security, which are recognized as best practices for safeguarding sensitive healthcare data.

In addition to using industry-standard security protocols, Arkstone further strengthens its data protection measures by implementing strict internal policies. One of the most critical aspects of Arkstone's approach is ensuring that its employees have minimal access to patient information. To this end, Arkstone enforces stringent access controls, ensuring that only authorized personnel can interact with data relevant to their role. The majority of employees are specifically restricted from viewing any personally identifiable patient information unless it is absolutely necessary for the performance of their duties. This practice helps to protect patient privacy and reduces the risk of accidental data breaches.

TRAINING AND VALIDATION OF DATA ENTERED

All data input into Arkstone undergoes a comprehensive training and validation process to ensure the accuracy and reliability of its decision-making. The first step in this process is to classify each data point as either "trained" or "untrained." Untrained data is flagged by the system for further training by a human expert in infectious disease, The system uses a combination of validation techniques to identify trained vs untrained data, including K fold validation, leave one out validation, random subsampling ensuring that new or previously unprocessed data is handled correctly. Since multiple data points are often transmitted simultaneously within each dataset, this process is executed across a variety of variables at once, streamlining the validation of complex datasets. These processes occur with every variable sent to Arkstone regardless of whether its trained or untrained data. However, the system does not solely rely on automation; a human expert is always involved to review and validate that the data is accurately classified as trained or untrained, ensuring the accuracy of the correlation before proceeding.

If a dataset contains data that has already been trained, the system performs an internal query to search for any other instances of the same exact variables in previous datasets. If a match is found, the system flags this for confirmation by a human in the loop to ensure that the data is consistent with previously trained information. In situations where an identical dataset has not been encountered previously, the system will identify the closest related match—drawing from the most similar data points within the training set. This closest match is then presented to a human expert for further review and validation. By using this process, Arkstone ensures that even novel data, which may not yet have a direct match in the system, is carefully reviewed and validated by a qualified professional before being integrated into the model. This layered approach, combining ML and expert oversight, guarantees the accuracy, relevance, and robustness of the data Arkstone uses to guide clinical decision-making.

SCALABILITY, REDUNDANCIES AND PROOFREADING

For scalability and long-term growth, automation is essential in ML systems. However, automation also presents significant challenges, including the risk of overfitting and potential errors. Arkstone has proactively addressed these issues through a multifaceted approach designed to maintain accuracy, consistency, and adaptability while ensuring human oversight remains central to the process.

First, automation within Arkstone's model is only permitted for datasets that have been trained consistently and in precisely the same manner at least twice by two separate experts in infectious disease. This assures minimal bias by one reviewer if any at all. In some cases, data points undergo up to seven rounds of training to ensure a robust foundation for automation which may include three or more reviewers. This stringent process guarantees that only highly validated data is subject to automated processing. Second, once data is fully trained and enters the automated phase, strict protocols prevent any confusion or intermixing with untrained data. The system's precision is demonstrated by its impressive F1 score of 1, reflecting its exceptional ability to differentiate between trained and untrained data with no false positives or negatives.

To further safeguard against errors and redundancy, Arkstone incorporates a dynamic re-training protocol even for fully trained data. At designated intervals, automated datasets are flagged for additional human review to confirm that no mistakes occurred during the initial rounds of training and to check for any newly emerging data that might necessitate updates. This mechanism ensures that even the most thoroughly trained data does not remain static but is periodically re-evaluated to reflect current clinical knowledge and evolving variables.

For example, a dataset that has been placed in automation after passing the training process more than twice will automatically revert to manual review after 15 successful automated cycles. At this point, a human expert will thoroughly examine the data to determine if it remains suitable for automation. If any modifications are required, the entire training process is restarted from the beginning. If no changes are needed, the dataset returns to automation but will again be pulled for review after another designated interval. This continuous feedback loop serves as a sophisticated proofreading mechanism, powered by clinical expertise, that reinforces data integrity and prevents complacency in automated processes.

This hybrid approach, combining automation with scheduled human oversight, provides a powerful balance between efficiency and safety. It ensures that the Arkstone system remains adaptive, error-free, and consistently aligned with the most current standards in infectious disease management.

ACCURACY OF THE CLINICAL RECOMMENDATIONS

The accuracy of Arkstone's clinical recommendations is grounded in a meticulous internal process designed to ensure that all guidance aligns with the most authoritative and current medical standards. As a core policy, Arkstone exclusively relies on primary sources of medical information. These include regulations and guidelines from esteemed organizations such as the U.S. Food and Drug Administration (FDA), the Centers for Disease Control and Prevention (CDC), the Infectious Diseases Society of America (IDSA), and other global health authorities. In addition to these foundational resources, Arkstone conducts comprehensive reviews of thousands of peer-reviewed research articles to ensure that its recommendations are firmly rooted in widely accepted, evidence-based guidelines.

This vast body of data undergoes thorough analysis by multiple infectious disease experts on the Arkstone team. These experts carefully evaluate relevant information and extract key points critical to determining appropriate treatment protocols. The source material for every clinical recommendation is meticulously documented within multiple layers of the Arkstone system, ensuring that clinicians have direct access to the original primary references. This transparency not only supports trust in the system's guidance but also empowers clinicians to verify the medical literature behind each recommendation.

Once a recommendation is derived from these trusted sources, it is subjected to multiple rounds of review by different members of the clinical team to verify its accuracy and appropriateness. Only after passing this rigorous vetting process is the information entered into Arkstone's database, where it is explicitly linked to specific pathogens and medical indications. When the system detects a pathogen in a patient's data, it can only generate a recommendation based on the precise, pre-approved treatment guidelines that have been manually programmed into the system. This ensures that all guidance provided by Arkstone is consistent with established medical protocols and thoroughly vetted expert consensus.

Arkstone's design as a non-predictive model is intentional and central to its mission of delivering dependable clinical recommendations. Unlike predictive models that rely on algorithmic patterns to generate outputs, Arkstone uses only predefined, validated responses based on medical guidelines. This approach guarantees that every recommendation adheres precisely to expert-reviewed standards, eliminating the risk of deviation from established medical practice. By using a fully transparent, manually curated system, Arkstone provides clinicians with the highest level of confidence in its clinical decision support, ensuring safe, evidence-based patient care.

In 2025, researchers at Arkstone Medical Solutions published an internal validation study of their machine learning–driven clinical decision support system (CDSS) for antimicrobial stewardship **(Frenkel A, *et al.* 2025)**. The system was tested in three ways: first, it demonstrated 100% accuracy (F1 = 1.0) in distinguishing 111 previously unseen variables from trained ones across nine training sessions using FDA-approved molecular diagnostic panels. Second, in an analysis of 1,401 real-world laboratory results drawn from 66 laboratories across 24 states and one international site, the model again achieved perfect precision and recall, correctly classifying 519 reports as fully trained and 644 as untrained, without any false positives or false negatives. Third, in a prospective review of 644 clinician-trained reports, independent infectious disease specialists found no major discrepancies with standard clinical guidelines and only 100 minor differences (15.5%), most involving alternative antibiotic choices, dosing variations, or questions about low-pathogenicity organisms. Collectively, these findings show that the system not only excels at differentiating training from novel data but also produces treatment recommendations that are highly consistent with expert clinical standards, underscoring its potential to enhance antimicrobial stewardship and reduce inappropriate antibiotic use.

REALTIME DATA

Arkstone offers real-time testing capabilities for data input, a feature accessible to any laboratory submitting information to the system. This functionality is particularly valuable during the initial integration phase, where laboratories can transmit comprehensive microbiology panels and immediately review the corresponding output to verify that it accurately reflects the input. Real-time data testing is not limited to the integration period—laboratories can conduct these tests at any time to support their own internal validation processes. This ensures that the data transmitted to Arkstone is processed correctly and that the outputs align with the original laboratory findings. Such testing helps build confidence in the accuracy and reliability of the clinical recommendations generated by Arkstone.

A further layer of quality assurance is built into Arkstone's framework through its strict policy of coupling all clinical recommendations with the original laboratory report. This integration provides clinicians with direct access to the unaltered lab results alongside Arkstone's expert guidance. By presenting both pieces of information together, Arkstone reinforces transparency and empowers healthcare professionals to independently verify the data. Clinicians can cross-reference the original findings with Arkstone's recommendations, adding a critical safeguard against potential discrepancies. This dual-reporting system enhances trust in the recommendations while simultaneously serving as a built-in mechanism for error detection and correction.

In addition, because Arkstone's infectious disease experts monitor data continuously and in real-time, the Arkstone team has a unique and proactive capability to detect unusual trends and anomalies within laboratory results. This vigilant oversight includes identifying patterns such as an unexpectedly high frequency of rare organisms being reported or combinations of pathogens that do not align with typical clinical presentations. Such discrepancies often raise red flags, suggesting potential issues that may stem from faulty laboratory equipment, manual input errors, sample contamination, or instances of colonization rather than true infection.

When these irregularities are observed, Arkstone's quality assurance process is immediately activated. As part of this rigorous system, the team investigates the data anomalies to ensure that the integrity of laboratory reporting is maintained. If suspicions are confirmed or further investigation is warranted, Arkstone's experts promptly reach out to the laboratory in question to communicate their findings. This proactive collaboration serves multiple purposes: it alerts the laboratory to possible technical or procedural errors that may require correction, supports the accuracy of future testing, and ensures that patient care is based on the most reliable and clinically appropriate information.

By integrating this layer of quality control, Arkstone not only enhances its commitment to delivering precise, evidence-based recommendations but also reinforces its partnership with laboratories in promoting diagnostic excellence. This continuous feedback loop between Arkstone's expert oversight and laboratory operations exemplifies a shared dedication to patient safety, operational efficiency, and the highest standards of clinical accuracy. Through these efforts, Arkstone ensures that both its recommendations and the foundational data they rely on meet the most stringent requirements for reliability and validity in infectious disease management.

RETROSPECTIVE RECOMMENDATION MODIFICATIONS

Because Arkstone only has access to a limited set of finite variables within a patient’s history—such as laboratory results, demographic information, Arkstone provides a unique feature called OneChoice Plus to enhance clinical decision-making. OneChoice Plus empowers clinicians by allowing them to manually input additional critical variables that may not have been included in the original lab submission. These variables can include newly identified allergies, resistance genes not reported by the laboratory, or key patient-specific factors such as renal and hepatic function. By incorporating these supplemental data points, clinicians can customize the recommendations to reflect the most accurate and comprehensive clinical picture for each individual patient.

This customizable feature ensures that clinicians remain in control of prescribing decisions, using Arkstone’s expert guidance as a powerful supplement rather than a directive. With OneChoice Plus, healthcare providers are equipped with the flexibility needed to address complex or evolving medical situations that require a more nuanced approach. For example, if a patient’s renal function is compromised, the clinician can adjust the medication selection or dosage within Arkstone’s framework to reflect these conditions, leading to safer and more effective treatment.

STRENGTHS, LIMITATIONS, AND WEAKNESSES

The Arkstone ML model demonstrates several notable strengths, including its rigorous validation process and commitment to data integrity. By employing human-in-the-loop (HITL) validation, Arkstone ensures that all inputs and outputs are carefully reviewed by infectious disease experts, minimizing errors and enhancing reliability. The system's adaptability to a wide range of variables and its ability to process diverse patient demographics across specialties further strengthen its generalizability and clinical utility. Additionally, Arkstone's emphasis on real-time data analysis and its transparency in linking recommendations to primary sources of medical information builds trust among clinicians. However, the model also has limitations. Its reliance on expert review, while beneficial for accuracy, can introduce delays and scalability challenges as the volume of data increases. In addition, although there are multiple rounds of human checks for each data set that is trained, human error is still possible. The system’s non-predictive nature, while designed to ensure safety and precision, limits its ability to proactively forecast outcomes or trends, which might hinder its competitiveness with predictive ML models in certain applications. Despite these limitations, Arkstone offers a robust framework for integrating ML into clinical decision-making while maintaining high standards of accuracy and ethical responsibility.

FUTURE RESEARCH

Future research on the Arkstone system should prioritize evaluating its impact on patient outcomes, particularly in terms of clinical accuracy, timeliness of treatment, and long-term health improvements. Studies assessing how the system contributes to reducing treatment errors, optimizing antimicrobial stewardship, and improving recovery rates would provide valuable insights into its real-world effectiveness. Also, studies comparing the recommendations generated by OneChoice with those made by human infectious disease experts could offer valuable insights into its effectiveness and applicability as well as insight into whether clinicians are likely to listen to ML models compared to human experts. Additionally, a detailed cost-benefit analysis is crucial to determine the economic viability of the system, including its potential to reduce healthcare costs by minimizing unnecessary treatments, hospitalizations, or laboratory inefficiencies. Research should also explore strategies to enhance the system’s integration into diverse clinical workflows, focusing on its interoperability with electronic health records (EHRs) and its adaptability to varying levels of technological infrastructure across healthcare settings. Addressing these areas will not only provide a clearer understanding of the system’s value but also identify opportunities to improve its accessibility, scalability, and overall applicability in both large healthcare organizations and smaller, resource-limited practices.

DISCUSSION

The Arkstone system represents a significant step forward in leveraging ML for clinical decision-making, particularly in infectious disease management. By combining advanced validation techniques with human oversight, it addresses many challenges faced by traditional ML models, such as data integrity, model drift, and ethical concerns. The inclusion of features like OneChoice Plus allows clinicians to customize recommendations based on unique patient variables, enhancing the system's applicability in diverse scenarios. However, its reliance on manual validation introduces scalability challenges, especially as data volumes increase. Furthermore, while the non-predictive nature ensures safety and precision, it limits the model's ability to proactively identify patterns or trends that could improve preventive care. Future enhancements should focus on integrating predictive capabilities while maintaining rigorous validation, improving interoperability with EHRs, and expanding research on its real-world impact on patient outcomes and cost efficiency. These advancements will further refine its role as a reliable tool for improving clinical decision-making.

CONCLUSION

The Arkstone model demonstrates a robust framework for integrating ML into clinical practice, offering significant strengths in accuracy, adaptability, and ethical responsibility. Its rigorous validation processes and transparency in recommendations build trust and ensure its alignment with current medical guidelines. With continued research into patient outcomes, cost analysis, and interoperability, Arkstone has the potential to become a critical component in the evolution of clinical decision support systems and antimicrobial stewardship improving healthcare delivery while maintaining a focus on evidence-based care.

BIBLIOGRAPHY

1. **Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... & Zieba, K.** (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.
2. **Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D.** (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
3. **Cabitza, F., Rasoini, R., & Gensini, G. F.** (2017). Unintended consequences of ML in medicine. JAMA, 318(6), 517-518.
4. **Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017).** Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.
5. **Feng, S., Li, X., Zhang, Y., & Du, Y.** (2019). Machine learning-based models for financial fraud detection. Journal of Financial Technology, 10(4), 24-32.
6. **Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A.** (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1-37.
7. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.
8. **Howard, A., & Shapiro, M.** (2018). IBM Watson: Disruptive innovation in healthcare. Health Systems, 7(1), 12-18.
9. **Kohavi, R.** (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137-1143.
10. **Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A.** (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.
11. **Mitchell, T. M.** (1997). Machine learning. McGraw-Hill Science/Engineering/Math.
12. **Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S.** (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
13. **Ross, C., & Swetlitz, I.** (2017). IBM Watson recommended "unsafe and incorrect" cancer treatments. STAT.
14. **Rudin, C.** (2019). Stop explaining black box ML models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

15. **Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P.** (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record-based clinical prediction. Journal of Biomedical Informatics, 83, 168-185.

16. **Strickland, E.** (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum, 56(4), 24-31.

17. **Voigt, P., & Von dem Bussche, A.** (2017). The EU General Data Protection Regulation (GDPR). Springer International Publishing.

18. **Zhou, Q., Peng, B., Yu, X., & Zhang, L.** (2018). Watson for Oncology: Evaluation of AI assistance in clinical oncology. The Lancet Oncology, 19(10), 1265-1274*.

19. **Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer.

20. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press.

21. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

22. **Jordan, M. I., & Mitchell, T. M.** (2015). Machine Learning: Trends, Perspectives, and Prospects. Science, 349(6245), 255-260.

23. **Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer.

24. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press.

25. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

26. **Jordan, M. I., & Mitchell, T. M.** (2015). Machine Learning: Trends, Perspectives, and Prospects. Science, 349(6245), 255-260.

27. **Herper, M.** (2018). IBM Watson's Health Struggles Show How Hard It Is to Use AI to Transform Health Care. Forbes.

28. **Jordan, M. I., & Mitchell, T. M.** (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

29. **Russell, S., & Norvig, P.** (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

30. **Cabitza, F., Rasoini, R., & Gensini, G. F.** (2017). Unintended consequences of ML in medicine. JAMA, 318(6), 517-518.

31. **Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A.** (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

32. **Rudin, C.** (2019). Stop explaining black box ML models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

33. **Yao, Y., Rosasco, L., & Caponnetto, A.** (2007). On early stopping in gradient descent learning. Constructive Approximation, 26(2), 289-315.

34. **Bergstra, J., & Bengio, Y.** (2012). Random search for hyper-parameter optimization. Journal of ML Research, 13, 281-305.

35. **Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer.

36. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press.

37. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

38. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

39. **Kohavi, R., & Provost, F.** (1998). Glossary of terms. Machine Learning, 30(2-3), 271-274.

40. **Molinaro, A. M., Simon, R., & Pfeiffer, R. M.** (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301-3307.

41. **Murphy, K. P.** (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

42. **Raschka, S., & Mirjalili, V.** (2019). Python Machine Learning: ML and Deep Learning with Python, scikit-learn, and TensorFlow (3rd ed.). Packt Publishing.

43. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

44. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

45. **Raschka, S., & Mirjalili, V.** (2019). Python Machine Learning: ML and Deep Learning with Python, scikit-learn, and TensorFlow (3rd ed.). Packt Publishing.

46. **Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer.

47. **Fawcett, T.** (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

48. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

49. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

50. **Raschka, S., & Mirjalili, V.** (2019). Python Machine Learning: ML and Deep Learning with Python, scikit-learn, and TensorFlow (3rd ed.). Packt Publishing.

51. **Bengio, Y., Courville, A., & Vincent, P.** (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

52. **Bergstra, J., & Bengio, Y.** (2012). Random search for hyper-parameter optimization. Journal of ML Research, 13, 281-305.

53. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press.

54. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

55. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

56. **Snoek, J., Larochelle, H., & Adams, R. P.** (2012). Practical bayesian optimization of ML algorithms. In Advances in Neural Information Processing Systems, 2951-2959.

57. **Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer.

58. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

59. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press.

60. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

61. **Raschka, S., & Mirjalili, V.** (2019). Python Machine Learning: ML and Deep Learning with Python, scikit-learn, and TensorFlow (3rd ed.). Packt Publishing.

62. **Pedregosa, F., *et al.*** (2011). Scikit-learn: ML in Python. Journal of ML Research, 12, 2825-2830.

63. **Abadi, M., *et al.*** (2015). TensorFlow: Large-Scale ML on Heterogeneous Systems. https://www.tensorflow.org/

64. **Chollet, F., *et al.*** (2015). Keras. https://keras.io/

65. **Bergstra, J., & Bengio, Y.** (2012). Random search for hyper-parameter optimization. Journal of ML Research, 13, 281-305.

66. **Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer.

67. **Fawcett, T.** (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

68. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

69. **Bergstra, J., & Bengio, Y.** (2012). Random search for hyper-parameter optimization. Journal of ML Research, 13, 281-305.

70. **Snoek, J., Larochelle, H., & Adams, R. P.** (2012). Practical bayesian optimization of ML algorithms. In Advances in Neural Information Processing Systems, 2951-2959.

71. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

72. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press.

73. **Raschka, S., & Mirjalili, V.** (2019). Python Machine Learning: ML and Deep Learning with Python, scikit-learn, and TensorFlow (3rd ed.). Packt Publishing.

74. **Abràmoff, M. D., *et al.*** (2018). "Pivotal trial of an autonomous AI-based diagnostic system for diabetic retinopathy in primary care offices." NPJ Digital Medicine.

75. **Esteva, A., *et al.*** (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature.

76. **Komorowski, M., *et al.*** (2018). "Artificial intelligence in intensive care medicine." Intensive Care Medicine.

77. **Lundberg, S. M., *et al.*** (2018). "Explainable AI for trees: From local explanations to global understanding." Nature Machine Intelligence.

78. **Mehrabi, N., *et al.*** (2021). "A survey on bias and fairness in machine learning." ACM Computing Surveys.

79. **McKinney, S. M., *et al.*** (2020). "International evaluation of an AI system for breast cancer screening." Nature.

80. **Niel, O., *et al.*** (2018). "Artificial intelligence can predict occurrence of acute kidney injury in critically ill patients using electronic health records." Nature Medicine.

81. **Rajkomar, A., *et al.*** (2018). "Scalable and accurate deep learning with electronic health records." NPJ Digital Medicine.

82. **Shickel, B., *et al.*** (2018). "Deep learning in healthcare: Review, opportunities and challenges." Briefings in Bioinformatics.

83. **Wynants, L., *et al.*** (2020). "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal." BMJ.

84. **Creswell, J.** (2019). Google's DeepMind faces a reckoning in health care. The New York Times.

85. **Hern, A.** (2017). Google's DeepMind and the NHS: what is the controversy about? The Guardian.

86. **Hern, A.** (2019). Google absorbs DeepMind Health. The Guardian.

87. **Kelion, L.** (2017). Google DeepMind 1.6m patient record deal 'inappropriate'. BBC News.

88. **Powles, J., & Hodson, H.** (2017). Google DeepMind and healthcare in an age of algorithms. Health and Technology.

89. **Wagner, K.** (2018). DeepMind Health joins Google Health. Vox.

90. **Farr, C.** (2018). Why IBM's Watson Health struggled to deliver big results in health care. CNBC.

91. **Herper, M.** (2018). IBM Watson's Health Struggles Show How Hard It Is to Use AI to Transform Health Care. Forbes.

92. **Howard, J., & Shapiro, M.** (2018). A Reality Check For IBM's AI Ambitions. IEEE Spectrum.

93. **Klein, A.** (2019). IBM Watson Health: A Hard Diagnosis. Healthcare IT News.

94. **Miliard, M.** (2020). IBM to sell parts of Watson Health to private equity firm. Healthcare IT News.

95. https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs

96. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

97. www.healthit.gov/sites/default/files/page/2023-04/NPRM_DSI_fact%20sheet-508.pdf

98. www.healthit.gov/sites/default/files/page/2024-01/DSI_HTI1%20Final%20Rule%20Presentation_508.pdf

99. **Frenkel A, Rendon A, Chavez-Lencinas C, Gomez De la Torre JC, MacDermott J, Gross C, Allman S, Lundblad S, Zavala I, Gross D, *et al.*** Internal Validation of a Machine Learning-Based CDSS for Antimicrobial Stewardship. Life. 2025; 15(7):1123. https://doi.org/10.3390/life15071123