Data Mining Tools for Knowledge Discovery

Special track at the 2nd International Conference on AI-Health, AIHealth 2025, March 9 – 13,
2025, Lisbon, Portugal, ISBN: 978-1-68558-247-0, URL: www.thinkmind.org

Les Sztandera
Thomas Jefferson University
Jefferson College of Life Sciences and Jefferson India Center
1020 Locus St., Philadelphia, PA 19107, USA
E-Mail: Les.Sztandera@Jefferson.edu

Abstract— **This paper summarizes four presentations in a session of the track "Data Mining Tools for Knowledge Discovery". The research work deals with the following key issues of this track: Unsupervised and Supervised Learning, Pattern Discovery, Association Rule Learning, Classification and Cluster Analysis, Applications of Data Mining in Life Sciences.**

Contributions in this track address research questions in Data Mining techniques to extract information, and subsequently patterns, in data not previously discovered. Research conducted at Thomas Jefferson University aims to aid in closing this gap of knowledge.

Keywords: pattern recognition, clustering, classification, life sciences

I. INTRODUCTION

Analyzing large amounts of data is a necessity. Data mining refers to extracting knowledge from large data repositories. Data mining techniques have found application in numerous scientific fields with the aim of uncovering previously unknown patterns and correlations, as well as predicting trends and behaviors. The goal of Data Mining Tools for Knowledge Discovery (DMTKD) is to present some the most common techniques and practices for knowledge discovery.

As a result, more and more researchers are motivated to develop algorithms for applications in Life Sciences. These aspects are highlighted in this special track.

II. SUBMISSIONS

The first paper on "Revolutionizing Prostate Cancer Diagnosis: An Integrated Approach for Gleason Grade Classification and Explainability" by Gavade et al. [1] is a state-of-the-art analysis of Prostate Cancer detection. Accurate grading of Prostate Cancer (PCa) is vital for effective treatment planning and prognosis. This study introduces an advanced framework for Gleason Grade (GG) classification, addressing challenges in accuracy, computational efficiency, and interpretability. Utilizing the SICAPv2 dataset, which contains annotated prostate biopsy Whole Slide Images (WSIs) graded from GG0 to GG5, the framework integrates cutting-edge machine learning and deep learning techniques. Feature extraction is performed using a custom-designed Variational Autoencoder (VAE) with a VGG16 backbone, chosen for its computational efficiency, while dimensionality reduction with Principal Component Analysis (PCA) optimally selects 50 features for classification. The classification pipeline combines machine learning models, including Support Vector Machines (SVM), logistic regression, and random forests, with custom Deep Neural Networks (DNNs). SVM with an Radial Basis Function (RBF) kernel achieved an accuracy of 84% following hyperparameter tuning, while a custom five-layer dense neural network incorporating dropout and batch normalization demonstrated superior performance with an accuracy of 94.6%. Explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP), gradient-weighted class activation mapping (GradCAM), and Local Interpretable Model-Agnostic Explanations (LIME), enhance model interpretability by providing insights into feature importance and aligning predictions with clinical expertise. This framework delivers a robust, scalable, and interpretable solution for automated GG classification, bridging the gap between advanced AI techniques and clinical application.

The second paper, "A Heavy Metals in Human Health and Pregnancy: How Data Analysis, Mining, and Modeling Present a Solution" by Frennborn et al. [2], deals with essential heavy metals, such as zinc, copper, iron, and manganese, play important roles in many biological processes. Other heavy metals, like lead, arsenic, cadmium, and mercury, can displace essential heavy metals and disrupt vital biological processes. Heavy metal exposure can be particularly detrimental to pregnant women and their developing fetuses; however, little is known about the combinatory impact that simultaneous exposure to multiple heavy metals may have on fetal development. Procuring a better understanding of how these metals influence fetal development is a critical first step to addressing this concerning lack of knowledge. There are numerous databases and datasets in existence storing data on heavy metal levels in maternal and fetal blood, and novel studies are being done to expand this collection of data. Data mining techniques present a tool that could be used to close this gap of knowledge by revealing patterns in data not previously discovered. Research at Thomas Jefferson University aims to aid in closing this gap of knowledge. In the study, data such as heavy metal blood concentrations, pregnancy complications, health outcomes, and demographic information will be collected from mothers and newborns. Data mining strategies will then be used to develop models capable of discovering data patterns. If this modeling is successful, such an approach can be utilized by healthcare providers in the future to assess patient risk and provide early intervention for at-risk pregnant patients.

The paper „Recent Advances in Machine Learning for Log File-Based PSQA for IMRT and VMAT" by De Jesus et al. [3] focuses on the critical need for a faster and more efficient approach to patient-specific quality assurance (PSQA) in radiation therapy. The accuracy of PSQA is crucial for the safety of radiation therapy, particularly with complex procedures like intensity-modulated radiation therapy (IMRT) and volumetric-arc radiation therapy (VMAT). Traditional phantombased methods, while effective, are time-consuming and fail to account for patient-specific variability and real-time treatment adjustment. To address these limitations, alternative strategies leveraging trajectory log files—automatically recorded during treatment—have emerged as promising tools for PSQA. In recent years, the application of machine learning and deep learning algorithms to trajectory log files has been increasingly studied in literature. These algorithms have shown notable progress in predicting PSQA outcomes and detecting errors, though further development is required before they can be fully integrated into clinical practice. By surveying key studies, the paper highlights the potential of algorithms such as support vector machines, treebased methods, and convolutional neural networks to enhance the efficiency and accuracy of log file-based PSQA. The findings underscore the promise of these techniques in replacing traditional methods while addressing current challenges to pave the way for clinical integration.

The fourth publication“Data Mining Techniques in Online Health Communities and Databases“ by Sweitzer and Mikkelson [4] focuses on online communities that have become a safe haven for patients that battle health conditions to garner common experiences. These platforms empower patients to act as healthcare consultants, in the form of reviewing drugs, devices, surgeries, and specific healthcare providers. Consequently, these online communities generate self-reported data on patient sentiment, which provide valuable insight to patient satisfaction connected to drug products and devices and adverse events associated to them. Data mining techniques can serve as potential tools to extract previously hidden information and patterns from online communities. Sentiment analysis serves as a classification model within supervised machine learning, where predictions are made and validated through associated characteristics. A classification algorithm is required to obtain maximum accuracy, though. The results generated by data mining could inform the pharmaceutical industry about in demand medical interventions according to patient needs. Consequently, the outputs of data mining could make a positive impact on current pre-clinical and clinical trials to streamline desired research according to patient sentiment and break the barrier between the bedside and benchtop. This review article highlights various data mining techniques that could be utilized to collect and transform data from online health communities, such as Facebook groups and Reddit. Section I introduces the concepts of online health communities, data mining, and sentiment analysis. Section II proposes different data mining methods that could help reduce the complexity of data obtained from these communities, including sentiment analysis, self-organizing maps, support vector machines, and the naïve Bayes classifier. Section III explores how data collected from such studies could be applied by the U.S. Food and Drug Administration (FDA) to identify adverse drug reactions and improve efficiency in drug production, such as vaccine development. Section IV addresses challenges related to data transformation and cleaning, proposing augmentations like web scraping and the Levenshtein distance method to address issues associated with data collection from online forums. Finally, Section V concludes the article, emphasizing the underutilization of data generated by online health communities and its potential to positively influence academic and institutional medical research.

CONCLUSION

Future challenges in Data Mining for Knowledge Discovery include maintaining data privacy and security, handling increasingly complex and diverse datasets, ensuring algorithm scalability, and addressing ethical concerns surrounding data usage and bias.

ACKNOWLEDGMENT

I could not be prouder of all the students participating in AI Health 2025 Conference after taking on the Data Science research challenges I had introduced in their doctoral courses at Jefferson College of Life Sciences. I am also very pleased with contributions from our partners in India, who collaborated with Jefferson on their research undertaking.

I, along with the students, Emma Frennborn, Khalil Rust, Kellin DeJesus-Michaels, Cali Sweitzer, and Cassandra Mikkelson, express gratitude and acknowledge travel support from Jefferson Global Health International Committee (GHIC) directed by Vice Provost for Global Affairs at Thomas Jefferson University, Dr. Richard Derman, who is responsible for overall coordination of global activities that include collaborative research, educational exchanges and clinical initiation. Kellin DeJesus-Michaels also acknowledges partial travel support from Jefferson Travel Alumni fund.

REFERENCES

[1] Anil B. Gavade, Rajendra B. Nerli, Shridhar C. Ghagane, and Les Sztandera, “Revolutionizing Prostate Cancer Diagnosis: An Integrated Approach for Gleason Grade Classification and Explainability“, in Special Track: Data Mining Tolls for Knowledge Discovery, AIHealth 2025, ISBN: 978-1-68558-247-0, www.thinkmind.org, 2025.

[2] Emma Frennborn, Khalil Rust, and Les Sztandera “Heavy Metals in Human Health and Pregnancy: How Data Analysis, Mining, and Modeling Present a Solution“, in Special Track: Data Mining Tolls for Knowledge Discovery, AIHealth 2025, ISBN: 978-1-68558-247-0, www.thinkmind.org, 2025.

[3] Kellin De Jesus, David Thomas, Leon Dunn, and Les Sztandera, “Recent Advances in Machine Learning for Log File-Based PSQA for IMRT and VMAT“, in Special Track: Data Mining Tolls for Knowledge Discovery, AIHealth 2025, ISBN: 978-1-68558-247-0, www.thinkmind.org, 2025.

[4] Cali Sweitzer and Cassandra Mikkelson, “Data Mining Techniques in Online Health Communities and Databases”, in Special Track: Data Mining Tolls for Knowledge Discovery, AIHealth 2025, ISBN: 978-1-68558-247-0, www.thinkmind.org, 2025.