Advancing Bioinformatic Research Through Artificial Intelligence: A Focus on Disease Prediction and Diagnosis

Sharan Gayathrinathan[1], *Frida M. Delgadillo[3], Brian I. Grajeda[2]*

[1] Department of Biological Sciences, University of Texas at El Paso, El Paso, TX 79968, USA; fmdelgadillo@miners.utep.edu (F.D.); sgayathrina@miners.utep.edu (S.G.); bigrajeda@utep.edu (B.I.G.)

[2] Border Biomedical Research Center, University of Texas at El Paso, El Paso, TX 79968, USA

[3] Environmental Science and Engineering Ph.D. Program, University of Texas at El Paso, El Paso, TX 79968, USA

* Correspondence: bigrajeda@utep.edu

† These authors contributed equally to this work

Received on: September 09, 2023 | Accepted on: January 2, 2024 | Published on: January 15, 2024

Abstract:

This review article outlines current methods in predictive modeling with various biological areas such as proteomic, genomics, multivariable analysis, and bioinformatics as a whole. The onset of the digital era has inadvertently led to the emergence of Artificial Intelligence (AI). The core foundation of many of these applications such as Neural Networks (NNs), and their evolved form, Deep Neural Networks (DNNs), have progressed to allow integration in biological and bioinformatic applications. This paper highlights the instrumental advancements developed by the integration of these technologies in disease prediction and prevention as well as biomarker development. Such advancements are permitted by the analysis of high-throughput proteomics and genomics data via machine learning algorithms. Moreover, the application of AI extends to various medical fields including cancer oncology, human aging research, diabetes, COVID-19, kidney diseases and cardiovascular diseases. This broad implementation range lends itself to the foundation for a new generation of advances in healthcare and medical research. Thus, the ongoing evolution of AI and machine learning algorithms can lead to the expansion of scientific investigation while simultaneously progressing the treatments and therapies currently available in the healthcare.

Key Word: *Artificial Intelligence, Bioinformatics, Biomarkers, Deep Learning, Machine Learning, Multivariable Analysis*

Introduction:

The technological revolution of the past few decades has ushered us into the digital age, with advancements resulting in the creation of AI. AI has a variety of applications in the fields of biology and bioinformatics, some of which include: profiling and predictions of disease states in

genomic, proteomic datasets, proteins structural modeling, multivariable modeling and drug discover to name a few. Further advancements of said applications could lead to substantial positive impacts in public health by disease prevention efforts using relevant biomarkers and current databases as well as the analysis of ongoing investigations.

When considering the implications and impacts of AI we must first understand the foundational architecture. Neural Networks (NNs) are generally the foundations of many AI applications1. These NNs can loosely be compared to the brain's architecture1. The NNs are comprised of many nodes which can be paralleled to neurons1. The arrangement of these nodes is generally set in three divisions: input, hidden and output. The connections have associated measures or weights that allow for the final result to be tailored and optimized. The model, through a process known as "training", adjusts based on the error rate and the disparity between predicted and actual outputs2. Training in this sense can fine tune the neural networks weights and better predict the outcomes2. This oversimplified explanation is what we consider machine learning (ML). When taken a step further the NNs can evolve into multi-layer complex networks known as Deep Neural Networks (DNNs)3 and by extension Deep Learning (DL). The DNNs complexity allows for the system to learn and help solve difficult problems that typical basic NNs would overlook. DNNs can also identify hierarchy of features, meaning that they allow for expansion into regions such as computer vision as seen in Krizhevsky et al.4 and natural language like the popular AI Generative Pre-trained Transformer 3 (GPT-3)5. DNNs applications are integrated in protein modeling such as the prediction tool AlphaFold6. They are also integrated genomic data sets like variant calling or gene expression analysis, the identification of biomarkers for biologically relevant predictions and a plethora of other applications.

Pertaining to bioinformatics, the addition of AI and ML has many beneficial integrations that we are continuing to explore. This review will spotlight some of the most impactful applications of these technologies, emphasizing their significant roles in advancing scientific inquiry and enhancing patient outcomes. Moreover, integrating AI algorithms is considered a valuable development, as it could potentially change the way we approach diagnosis and treatments of diseases. This review article will encompass: AI in genomics, prediction and prevention, biomarker development and protein structure prediction.

Artificial Intelligence in Genomics:

Preprocessing is crucial in AI algorithms as it standardizes data for downstream use. This process addresses challenges such as noisy and/or missing data, as well as dimensionality issues, with the support of recent software and tools7. Traditional ML methods such as linear regression and support vector machines (SVMs) use statistical models to learn data patterns, providing key applications in biomedical genomic research, such as the coronavirus disease 2019 (COVID-19) and cancer8. Deep neural networks (DNNs), which have multiple hidden layers, along with convolutional neural networks (CNNs) — a specialized class of deep learning models for visual data analysis — are equipped to tackle genomics challenges through their unique architectures. Meanwhile, recurrent neural networks (RNNs), designed to identify patterns within sequential data, play a critical role in analyzing time-based sequencing datasets9. COVID-19 brought many recent changes to the field due to the extent of its impact and the necessity of accelerated research to mitigate the spread. An effective response to the COVID-19 epidemic required rapid genomic segmentation. Randhawa GS used a non-aligned, machine-learning method that rapidly and accurately segments the COVID-19 genome, confirming its origins in the betacoronavirus10. The strength of the method lies in the analysis of raw DNA sequences without needing genome

annotation, highlighting the complexity of coronavirus evolution and the ability of ML algorithms to perform genomic analysis in a timely manner to combat large scale virus outbreaks. The employment of AI algorithms increases accuracy and precision over manual analysis methods, and at the same time reduces human error. By combining genomic data with current health records and environmental information, AI-aided procedures provide a holistic health perspective that could advance genomics, decision-making, disease understanding, medical innovation, and personalized health care as well as disease prediction and prevention methodologies.

AI in Disease Prediction and Prevention:

There are many models and tools that can help predict and gauge the progression of diseases. Each of these tools has a foundational core based on NNs. Ensemble Learning is a ML technique that takes multiple categories to make more precise predictions in contrast to simply using one classifier11. The methodology has been used in a variety of disease states such as diabetes12, skin disease13, kidney disease14, liver disease15, and heart conditions16. The methods to Ensemble Learning have a variety of approaches that include bagging, boosting, stacking, and voting, each of which can play a critical role in the design of the model11. Briefly, these approaches each have distinctive functionalities: bagging (Bootstrap Aggregating) improves the stability and accuracy of machine learning algorithms12, boosting reduces bias and builds strong predictive models17, voting is used when combining conceptually different machine learning classifiers to distinguish the optimal one13, and finally stacking involves combining the predictions from multiple models to train a new model18. The combination of these approaches allows for a refined and accurate prediction. Graph Neural Networks (GNNs) is another interesting technique that uses graphs as an input data for predictions. In contrast to data vectors, graphs can convey complex data structures that numerical datasets are sometime difficult to extrapolate from. These GNN models have

potential to aid not only in disease prediction but help in medical diagnosis and treatment19,20. GNN's applications extend to prediction of protein-protein interactions21, prediction of drug interactions with proteins22, and relationship characterization of brain imaging23.

These ML models have been integrated with patient data in order to structure methods for prediction and prevention assessments. There are various studies in extrapolating this information for predictive analysis as well as various ML algorithms. In Khalid et al. we see a variety of algorithms implemented like Naïve Bayes, decision tree, K-nearest neighbor, random forest, support vector machine, Linear Discriminant Analysis (LDA), Gradient Boosting (GB), and neural network24. These algorithms can be applied to determine if a patient has Chronic Kidney Disease or not to a 100% accuracy24. In Arumugam et al., the forecasting of heart disease and diabetes was improved using ML and fine-tuned decision tree models which outperformed the naïve Bayes and support vector machine models25. These models used multiple variables to predict various diseases, thus demonstrating MLs ability to handle complex multivariable data in a healthcare setting25. In You et al., prediction models for Cardiovascular diseases (CVD) were obtained by using known empirical clinical knowledge and a list of comprehensive variables26. These variables or predictors were selected using ML and the research group was able to develop a novel CVD risk prediction model26. Additionally, by implementing the model created by You et al., intervention of high-risked CVD partients will help aid in the preventive clinical decisions26.

Additionally, another aspect of multivariable ML applications include the integration of various omic analysis. Capturing the intricate interplay within biological systems necessitates the integration of genomic, transcriptomic, and proteomic information. Databases such as 'LinkedOmics'27 offer an extensive repository of cancer-related omics data, which is invaluable for training predictive models. Take central nervous system tumors, for instance, where multi-

omic analyses have revealed predictive markers of tumor progression28. Machine learning excels in sifting through these vast and complex datasets, bringing to light new facets of tumor biology that have significant implications for diagnosis and prognosis in oncology29. Beyond aiding multivariable analyses for disease prediction, these machine learning models also set the stage for breakthroughs in biomarker discovery.

AI in Biomarker Development:

One additional implementation of AI lies in the development of biological markers in efforts to improve healthcare and make advances in medical research. Biological markers, also known as biomarkers, is a broad term that encapsulates the objective signs of a disease or condition that can be accurately measured30,31. Biomarker assessments draw from clinical data, which categorize molecular markers found in patient samples like blood and bodily fluids31,32. Additionally, machine learning algorithms process extensive genetic and proteomic data, further supporting diagnostic and monitoring efforts. In relation to diseases, healthcare workers can utilize biomarkers to detect the presence of diseases and monitor their progression by providing insights to its severity. They can be advantageous in personalized medicine, to match patients to the treatments that is best suited to complement their genetic makeup and to assess the individual's receptivity to particular treatments33. For example, Rezayi et al. reviewed AI techniques and their effectiveness in neoplasm precision medicine33. It was identified that 34 papers containing patient genomic, somatic mutation, phenotype, and proteomics with drug-response data was used as input in AI methods33. Additionally 16 papers using AI approaches looked at drug responses, a functional category for personalized treatment33.

One potential avenue for this development could lie in the application of AI to the growing high throughput proteomic data sets obtained by MS-based proteomics34. DL can analyze said MS

proteomic data and has now become a vital part of the data generation pipeline for biomarker discovery35. In a recent publication from Nakayasu et al., through a ML analysis, they identified protein panels capable of predicting the emergence of persistent autoantibodies and Type 1 diabetes (T1D) even six months before the autoimmune response appeared36. The authors advocate for evaluating these predictive protein panels in ongoing human cohort studies for better prognostics and therapeutics development concerning autoimmunity and T1D36. Despite these recent developments, biomarkers are novel, and their development requires various efforts. One study by Xiao et al., highlights the benefit and necessity of biomarker development in cancer oncology in relation to screening, diagnosis, and therapeutics34. Additionally, AI intervention could diminish the amount of time spent in cancer identification, by advancing precision oncology via biomarker evaluation37. Apart from established diseases, the lack of biomarkers is evident in novel applications such as the development of anti-aging remedies. The study by Putin et al selected 21 DNNS to predict human chronological age using blood samples from routine health exams in the hopes of facilitating the tracking of biomarkers38. This led researchers to develop an online system to evaluate the performance of the predictors, which could potentially lead to the expansion of DNN training for the analysis of different types of biological data38. Even though monitoring concentration of biomarkers can help in elucidating disease conditions, some more thorough structural approaches are needed to truly understand the functionality of specific biomolecules.

AI in Protein Structure Prediction:

The protein structure prediction software AlphaFold2 has allowed the identification of over 200 million protein structures39. Of the structures generated they have be complemented with cryogenic electron microscopy (cryo-EM) to help elucidate critical structural biology tasks, such

as functional classification, variant effects, binding site prediction and modeling into new experimental data40. Alphafold2 has helped in a variety of applications including: identification of nuclear pore complex proteins41, characterization of molecular mechanisms for the activation of gametogenesis in malaria parasites42 and elucidation of CCR4–NOT transcription complex subunit 9, a key player in mRNA degradation43.

As of July 2020, AlphaFold2, previously known as AlphaFold6, is currently the best method for protein structural predictions44. The model was entered in the CASP14 assessment and had a significantly better accuracy compared to other models44. The evaluation conducted by CASP occurs every two years, utilizing newly resolved structures that have not been registered in the PDB or publicly revealed, ensuring a blind test scenario for the methods partaking in the assessment. This evaluation has historically stood as the benchmark for gauging the precision of structure prediction endeavors45,46. AlphaFold2 strength and accuracy in protein structure prediction come from novel neural network architectures and refined training procedures, while adhering to the evolutionary and geometric principles of proteins. It employs a unique architecture to jointly embed Multiple Sequence Alignments (MSAs) and pairwise features, bolstering end-to-end structure prediction. The equivariant attention architecture and a structure module work in tandem to elucidate precise 3D coordinates of protein residues from amino acid sequences. The "Evoformer" neural network block processes inputs and connects information about spatial and evolutionary relationships within proteins. Furthermore, the iterative refinement strategy termed 'recycling' significantly enhances prediction accuracy with a slight extension in training time. The structured methodology, iterative refinements, and the innovative architectures together encapsulate AlphaFold2’s strategy in decoding the intricate 3D structure of proteins44. Though

great strides have been made to accurately predict these models, we are far from accurately determining the nuances in protein configuration when other interacting partners are present47.

Discussion:

AI's integration as a tool in scientific research brings forth a multitude of transformative possibilities. Notably AI can aid researchers by improving their efficiency in analyzing large-scale datasets that are often insurmountable due to the near impossibility of single-handed human interpretation. Among these, the applications relating to disease prediction, biomarker development and proteomic/genomic analysis will serve not only to complement and accelerate current research projects but also to improve the predictive insights we gain from simulations and models. Currently, many models have been constructed to predict or identify diseases and changes in metabolism such as: kidney diseases, cardiovascular diseases, cancer, COVID-19, diabetes, and aging to name a few.

By examining complex data and detecting subtle patterns, AI can simultaneously lead to effective disease prediction, management and treatment, with the compiled information, enhancing biomarker identification and development. AI significantly refines proteomic and genomic data analysis, illuminating the complex genetic and protein dynamics fundamental to biological processes. Furthermore, as AI and ML algorithms continue to evolve, they will open up new avenues of scientific investigation considered to be distantly unobtainable by today's standards. By predicting outcomes, simulating experiments, and optimizing processes with an unprecedented level of sophistication and efficiency, AI expands the horizons of scientific exploration. Multiple algorithms have been implemented in conjunction with AI such as Naïve Bayes, decision tree, K-nearest neighbor, random forest, SVM, LDA, GB, and neural networks. Techniques like Ensemble Learning have also included an array of tools from bagging, boosting, stacking, and voting. GNNs

extend to prediction of protein-protein interactions, drug interactions and an array of real-world clinical applications. DNNs and by extension CNNs and RNNs aid in visual data analysis and sequential data for refined parsing. AlphaFold2 has unlocked an array of structural protein information that is invaluable to clinical applications. All these tools in conjunction and complementation to AI shape the methodologies that allow for breakthroughs in scientific inquiry and clinical progression.

Though much progress on protein modeling and scaffolding has been accomplished, there needs to be a push for translational science to allow for health care professionals to make better decisions for patients. The same idea can be mentioned for predictive models. There are many datasets and repositories as well as models that have been designed to predict and prevent disease state. Although the applications of AI in healthcare are not yet broadly utilized, the models trained on pre-existing datasets often embed inherent biases and disparities. If not reduced appropriately, AI algorithms could perpetuate existing inequities and data gaps in the field, like the healthcare sector48. Thus, it is important to use representative data sets and to robustly address potential biases in algorithmic development. While novel, the implementation of AI is simply an extension of the digital age that led to the scientific discoveries that have aided humankind of previous decades and will continue to allow for the progression of such discoveries in the distant future.

Author Contributions: Conceptualization, B.I.G.; writing—original draft preparation, F.M.D., S.G. and B.I.G.; writing—review and editing, F.M.D., S.G., B.I.G.; validation, B.I.G; supervision, B.I.G.; All authors have read and agreed to the published version of the manuscript.

Acknowledgments: A special thanks to Charlie T. Gupta

Conflicts of Interest: The authors declare no conflict of interest.

References:

1. Bengio, Y., Lecun, Y. & Hinton, G. Deep learning for AI. Commun. ACM 64, 58–65 (2021).

2. Bishop, C. M. Pattern Recognition and Machine Learning. (Springer New York, 2006).

3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).

4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90 (2017).

5. Brown, T. B. et al. Language Models are Few-Shot Learners. arXiv [cs.CL] (2020).

6. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020).

7. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. & Herrera, F. Big data preprocessing: methods and prospects. Big Data Anal. 1, (2016).

8. Jenni, A. M. & Sidey-Gibbons, C. J. Machine learning in medicine: a practical introduction. BMC medical research methodology 19, 1–18 (2019).

9. Zafar, A. et al. A comparison of pooling methods for convolutional neural networks. Appl. Sci. (Basel) 12, 8643 (2022).

10. Randhawa, G. S. et al. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS One 15, e0232391 (2020).

11. Mahajan, P., Uddin, S., Hajati, F. & Moni, M. A. Ensemble learning for disease prediction: A review. Healthcare (Basel) 11, 1808 (2023).

12. Ganie, S. M. & Malik, M. B. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. Healthcare Analytics 2, 100092 (2022).

13. Shehzad, K. et al. A deep-ensemble-learning-based approach for skin cancer diagnosis. Electronics (Basel) 12, 1342 (2023).

14. Eroğlu, K. & Palabaş, T. The impact on the classification performance of the combined use of different classification methods and different ensemble algorithms in chronic kidney disease detection. National Conference on Electrical, Electronics and Biomedical Engineering 512–516 (2016).

15. Majzoobi, M. M., Namdar, S., Najafi-Vosough, R., Hajilouei, A. A. & Mahjub, H. Prediction of hepatitis disease using ensemble learning methods. (2022) doi:10.15167/2421-4248/JPMH2022.63.3.2515.

16. Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L. & Javed, T. Cardiovascular Disease Detection using Ensemble Learning. Comput. Intell. Neurosci. 2022, 1–9 (2022).

17. Singh, S. & Gupta, S. Prediction of diabetes using ensemble learning model. in Advances in Intelligent Systems and Computing 39–59 (Springer Singapore, 2021).

18. Laila, U. e., Mahboob, K., Khan, A. W., Khan, F. & Taekeun, W. An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. Sensors (Basel) 22, 5247 (2022).

19. Lu, H. & Uddin, S. Disease prediction using graph machine learning based on electronic health data: A review of approaches and trends. Healthcare (Basel) 11, 1031 (2023).

20. Kim, S. Y. GNN-surv: Discrete-time survival prediction using Graph Neural Networks. Bioengineering (Basel) 10, 1046 (2023).

21. Zhang, X.-M., Liang, L., Liu, L. & Tang, M.-J. Graph neural networks and their current applications in bioinformatics. Front. Genet. 12, (2021).

22. Tran, H. N. T., Thomas, J. J. & Ahamed Hassain Malim, N. H. DeepNC: a framework for drug-target interaction prediction with graph neural networks. PeerJ 10, e13163 (2022).

23. Zhang, S. et al. The combination of a graph neural network technique and brain imaging to diagnose neurological disorders: A review and outlook. Brain Sci. 13, 1462 (2023).

24. Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G. & Shuaib Qureshi, M. Machine learning hybrid model for the prediction of chronic kidney disease. Comput. Intell. Neurosci. 2023, 1–14 (2023).

25. Arumugam, K. et al. Multiple disease prediction using Machine learning algorithms. Mater. Today 80, 3682–3685 (2023).

26. You, J. et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke Vasc. Neurol. svn-2023-002332 (2023).

27. Vasaikar, S. V., Straub, P., Wang, J. & Zhang, B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 46, D956–D963 (2018).

28. Takahashi, S. et al. A New Era of neuro-oncology research pioneered by multi-omics analysis and machine learning. Biomolecules 11, 565 (2021).

29. Hunter, B., Hindocha, S. & Lee, R. W. The role of artificial intelligence in early cancer diagnosis. Cancers (Basel) 14, 1524 (2022).

30. Mandal, A. & Sally Robertson, B. S. What is a biomarker? News-medical.net https://www.news-medical.net/health/What-is-a-Biomarker.aspx (2010).

31. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin. Pharmacol. Ther. 69, 89–95 (2001).

32. Hirsch, M. S. & Watkins, J. A comprehensive review of biomarker use in the gynecologic tract including differential diagnoses and diagnostic pitfalls. Adv. Anat. Pathol. 27, 164–192 (2020).

33. Rezayi, S., R Niakan Kalhori, S. & Saeedi, S. Effectiveness of artificial intelligence for personalized medicine in neoplasms: A systematic review. Biomed Res. Int. 2022, 1–34 (2022).

34. Xiao, Q. et al. High-throughput proteomics and AI for cancer biomarker discovery. Adv. Drug Deliv. Rev. 176, 113844 (2021).

35. Mann, M., Kumar, C., Zeng, W.-F. & Strauss, M. T. Artificial intelligence for proteomics and biomarker discovery. Cell Syst. 12, 759–770 (2021).

36. Nakayasu, E. S. et al. Plasma protein biomarkers predict the development of persistent autoantibodies and type 1 diabetes 6 months prior to the onset of autoimmunity. Cell Rep. Med. 4, 101093 (2023).

37. Fitzgerald, J. et al. Future of biomarker evaluation in the realm of artificial intelligence algorithms: application in improved therapeutic stratification of patients with breast and prostate cancer. J. Clin. Pathol. 74, 429–434 (2021).

38. Putin, E. et al. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging (Albany NY) 8, 1021–1033 (2016).

39. Callaway, E. “The entire protein universe”: AI predicts shape of nearly every known protein. Nature 608, 15–16 (2022).

40. Akdel, M. et al. A structural biology community assessment of AlphaFold2 applications. Nat. Struct. Mol. Biol. 29, 1056–1067 (2022).

41. Fontana, P. et al. Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. Science 376, eabm9326 (2022).

42. Zhu, C. et al. Characterizing the specific recognition of xanthurenic acid by GEP1 and GEP1-GCα interactions in cGMP signaling pathway in gametogenesis of malaria parasites. Int. J. Mol. Sci. 24, 2561 (2023).

43. Pavanello, L., Hall, M. & Winkler, G. S. Regulation of eukaryotic mRNA deadenylation and degradation by the Ccr4-Not complex. Front. Cell Dev. Biol. 11, (2023).

44. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

45. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. Proteins 23, ii–v (1995).

46. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins 87, 1011–1020 (2019).

47. Saey, T. H. Has AlphaFold actually solved biology's protein-folding problem? Science News Magazine https://www.sciencenews.org/article/alphafold-ai-protein-structure-folding-prediction (2022).

48. Gurupur, V. & Wan, T. T. H. Inherent bias in artificial intelligence-based decision support systems for healthcare. Medicina (Kaunas) 56, 141 (2020).