Elevating Rapid Learners' Skills through WiDS Datathon Participation

[1]Sujatha C, [2]Padmashree Desai, [3]Meena S. M, [4]P. G, Sunita Hiremath,
[5]Sneha Varur, [6]Neha T, [7]Vijayalakshmi M
School of Computer Science and Engineering, KLE Technological University, Hubli
[1] sujata_c@kletech.ac.in, [2] padmashri@kletech.ac.in, [3] msm@klectech.ac.in,
[4] sneha.varur@kletech.ac.in, [5] pgshiremath@kletech.ac.in ,[6] neha_ip@kletech.ac.in,
[7] viju11@kletech.ac.in

***Abstract*— In this paper, we explore the integration of the Women in Data Science (WiDS) Challenge into an Exploratory Data Analysis (EDA) course as a strategic approach to elevate the skills of rapid learners. The WiDS Challenge, a globally recognized data science competition, provides students with a unique opportunity to apply theoretical knowledge to real-world problems in a collaborative and competitive environment. The WiDS Datathon offers a real-world data science global challenge that allows students to apply EDA techniques in a practical, competitive setting. Data science is crucial in education. A course on this subject aids student in comprehending data, performing analysis, building models, and drawing inferences from results. These skills are applicable in real-world scenarios such as healthcare, climate change, agriculture, and finance. By embedding the challenge within the EDA curriculum, students are exposed to hands-on data analysis, model development, and problem-solving exercises that reinforce classroom concepts. The approach fosters rapid skill development, critical thinking, and teamwork, while also promoting diversity and inclusion in the field of data science. The results indicate that the Datathon significantly boosts rapid learners' analytical capabilities and practical knowledge, demonstrating its value as an effective and engaging pedagogical tool for advanced EDA students. This study underscores the potential of project-based learning in enhancing the educational experience and preparing students for real-world data challenges. Students earned the credits because it was integrated into the course, which contributed to the teams' strong performance. As a result, the top three teams from our university improved their global rankings to 10th, 14th, and 47th, respectively, compared to the previous year.**

***Keywords*— Data preparation; Data analysis; project-based learning, course project, WiDS Datathon.**

ICTIEE Track*: Assessment of effective teaching*

ICTIEE Sub-Track: Assessment for Learning: Empowering Students through effective Assessment Practices

I. INTRODUCTION

Exploratory Data Analysis (EDA) is a critical initial step in the data analysis process, where analysts explore datasets to uncover underlying patterns, spot anomalies, test hypotheses, and check assumptions through visual and quantitative methods. It involves summarizing the main characteristics of the data, often with visualizations, to gain a better understanding of its structure and relationships before formal modeling. EDA helps in identifying trends, correlations, and potential outliers, enabling data scientists to make informed decisions about which techniques to use in the subsequent stages of analysis. The insights gained during EDA guide the selection of appropriate statistical models and ensure that the data is well-prepared for more complex analyses.

The main objective was to introduce students early to global challenges, thereby enhancing the skills of quick learners by offering them opportunities. Since the EDA course primarily focuses on real-world data analysis, our goal is to bridge theory with practice by adopting a strategic approach for the course project. Each year, the WiDS community hosts a Kaggle competition to empower women in data science. Previously, our senior students participated in this challenge as an additional co-curricular activity alongside their regular semester courses, resulting in minimal participation and satisfactory performance. From the literature, we observe that most papers primarily emphasize on the EDA concept teaching and usage of tools. Additionally, we also prioritize the skill development of rapid learners.

In 2024, a university edition datathon was introduced, allowing us to integrate the datathon into the curriculum and assign credits to students. This change motivated us to provide students with hands-on experience and prepare them for real-world datathon challenges. While data analysis has always been part of the course project, participating in global challenges requires students to also engage in model building, which demands rapid learning skills. As a result, we identified a group of fast learners and formed teams to participate in the global challenge, with the goal of enhancing their skills and preparing them for high-level competition. However, the coverage of topics and lab programming was challenging to align with the competition. To address this, a workshop on model building and performance evaluation was conducted well before the competition deadline, providing hands-on experience.

II. LITERATURE REVIEW

Exploratory Data Analysis (EDA), introduced by John Tukey in 1977, is an approach to analyzing datasets that emphasizes summarizing their main characteristics, often through visual methods. EDA is a vital step in the data analysis process because it enables data scientists to understand the data, identify patterns, detect anomalies, formulate hypotheses, and gain insights before moving on to formal modeling. The literature highlights the importance of EDA in the initial phase of data analysis to ensure a robust and effective downstream process.

Numerous studies and textbooks, such as "Exploratory Data Analysis with R" by Roger D. Peng, focus on the techniques and tools used in EDA, including summary statistics, data visualization, and graphical analysis. Tools like R, Python, and their respective libraries (e.g., ggplot2, matplotlib, seaborn) are extensively utilized in teaching EDA. These resources often emphasize hands-on learning with real-world datasets, which can be aligned with Sustainable Development Goals (SDG) data for contextual learning.

The works in Beckman, M. D., & Cook, D. (2021) to Ho, S. H., & Tai, J. H. (2020), discuss modern approaches and tools for teaching EDA and data visualization, advanced techniques for handling high-dimensional data, and the challenges and trade-offs. They also provide an in-depth analysis of interactions and evaluate the use of map-based decision support tools for EDA in environmental contexts. Meanwhile, the works carried out in Comings, D. E. (2001) to El Malhany, N et al (2015) explore neurodevelopmental disorders such as Attention-Deficit/Hyperactivity Disorder (ADHD) and Tourette Syndrome (TS), which often co-occur, and present research findings on these conditions.

The works related to breast cancer were explored to understand the domain and solution approaches using AI-ML. The works in Chen, T. B., et al. (2021) to Fisher, R. T., et al. (2024). highlight how AI and ML are increasingly being integrated into breast cancer research, from enhancing diagnostic accuracy to improving personalized treatment strategies.

III. METHODOLOGY

The Exploratory Data Analysis (EDA) course is available to B.E. Computer Science and Engineering students in their 4th semester, carrying 2 credits for theory and 2 for lab work (2-0-2). It consists of two hours of theoretical instruction and four hours of laboratory work each week. The course employs a project-based learning approach, with concepts explored and applied in the lab using Python notebooks.

Given that the EDA course focuses heavily on real-world data analysis, our aim is to integrate theory with practice through a strategic course project, as illustrated in Fig 1. The theory includes EDA concepts such as, Descriptive Data Analysis and Visualization, Data Pre-Preprocessing, Time-series analysis, supervised and unsupervised learning. Python programming is introduced towards implementing the concepts, data analysis and visualization techniques.

The WiDS Datathon encourages women worldwide to hone their data science skills, creating a supportive environment for women to connect with others in their community who share their interests. Data scientists of all levels are invited to participate in the datathon, including beginners.

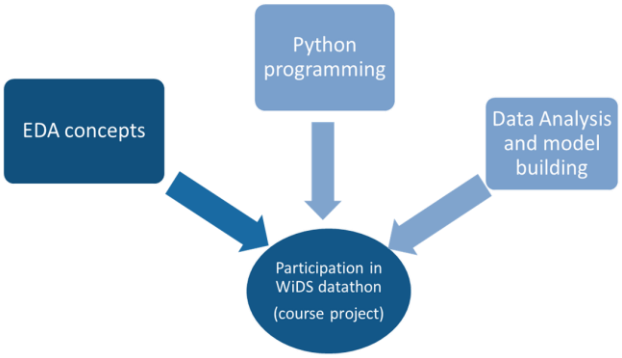

Python programming
EDA concepts
Data Analysis and model building
Participation in WiDS datathon (course project)

Fig 1: Strategic approach for the rapid learners.

Description of WiDS Datathon Challenge 2024: The challenge was to predict the duration of time it takes for patients to receive metastatic cancer diagnosis.

Dataset Description: Contained 19k records and 150 attributes such as Patients and their characteristics (age, race, BMI, zip code), Diagnosis and treatment information (breast cancer diagnosis code, metastatic cancer diagnosis code, metastatic cancer treatments etc.), Geo (zip-code level), Demographic data (income, education, rent, race, poverty etc), and climate data. The performance of the model was evaluated using the metric Root Mean Squared Error (RMSE). Based on this metric the ranking was assigned to the teams.

While data analysis has always been part of the course project, participating in global challenges requires students to also engage in model building, which demands rapid learning skills. To meet this requirement, we followed a process to elevate the skill of rapid learners as shown in Fig 2.

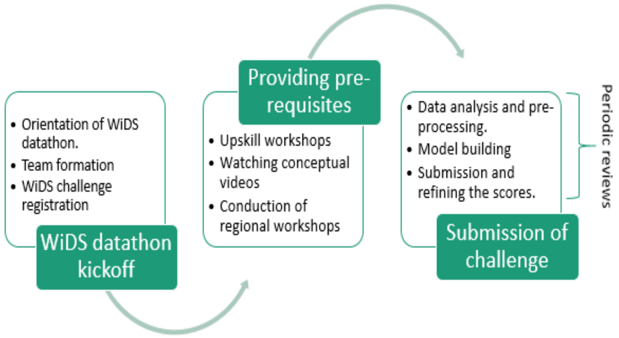

Orientation of WiDS datathon.
Team formation
WiDS challenge registration
WiDS datathon kickoff
Providing pre-requisites
Upskill workshops
Watching conceptual videos
Conduction of regional workshops
Data analysis and pre-processing.
Model building
Submission and refining the scores.
Periodic reviews
Submission of challenge

Fig 2: Process of WiDS participation

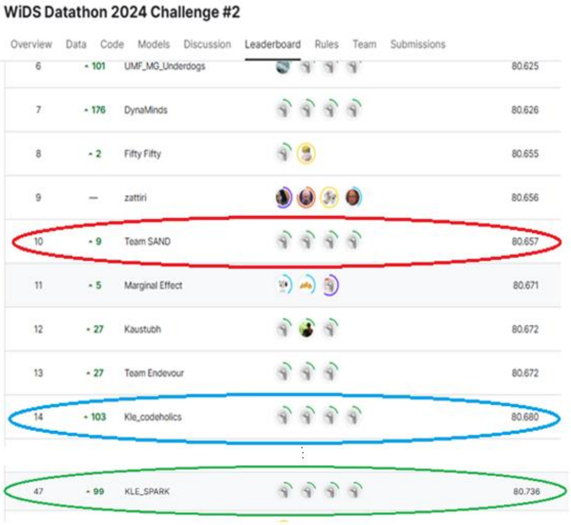
The aim was to make the students ready for the WiDS challenge, participate and compete globally. Initially the orientation of the WiDS datathon was introduced, then the team formation was made of size 4 where 50% are women as per the constraint. The team registered for the challenge on Kaggle platform. The task of the challenge was complex, so the necessary pre-requisites need to be met, so the regional workshop was conducted to assist the participants to enhance the required data science skillsets and participate effectively. Data science experts engaged in the sessions during this workshop and provided insights into analytical solutions for real world problems. In addition to this, students attended the upskill workshop organized by other regional events. The course teachers periodically reviewed the progress of the teams phase wise and assisted to refine the results. The senior students who had previously participated in the WiDS challenge shared the experience to the juniors and provided the tips how to attempt the datathon starting from data pre-processing to model building and submission of the solution on the Kaggle platform.

IV. RESULTS AND DISCUSSIONS

After the attending the workshops the WiDS teams were confident to solve the datathon task. The teams made the submissions periodically and refined the scores based on their rankings. The RMSE metric was used to obtain the ranks by the host. We made a comparative study with the last two years' teams and observed the integration and the credits earned via the datathon supported to perform well in the year 2024 as shown in Table I. We observe that the Datathon integration has not only increased the number of participation teams but also the global ranking has been improved to 10, 14 and 47 respectively for the top teams within our university. The leaderboard of WiDS datathon is shown in Fig 3 for the top 3 teams global ranking (10,14, 47) within our university.

TABLE I
COMPARISON OF THE PERFORMANCE WITH PREVIOUS YEARS.

	Sem	Datathon integrated with curriculum	#teams participated globally	# teams participated from KLE Tech	Global ranking (top 3 within our university)
2022	VI	No	512	8	126, 129 and 137
2023	VI	No	697	16	39, 64, 68
2024	IV	Yes	542	26	**10, 14**, 47



WiDS Datathon 2024 Challenge #2

6	▴ 101	UMF_MG_Underdogs	80.625
7	▴ 176	DynaMinds	80.626
8	▴ 2	Fifty Fifty	80.655
9	—	zattiri	80.656
10	▴ 9	Team SAND	80.657
11	▴ 5	Marginal Effect	80.671
12	▴ 27	Kaustubh	80.672
13	▴ 27	Team Endevour	80.672
14	▴ 103	Kle_codeholics	80.680
47	▴ 99	KLE_SPARK	80.736

Fig 3: Leaderboard of WiDS datathon

We also analyzed the distribution of range of ranks from our university and see that four teams have scored within 100, 9 between 190 to 280 and more teams fall in the 290 to 370 as shown in Fig 4. Since the students belong to lower semester, the performance can be considered as satisfactory.

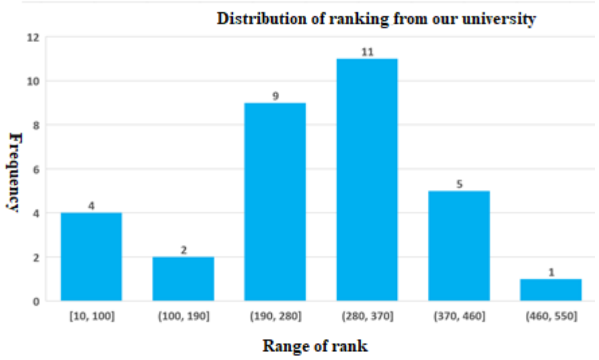


Fig 4: Distribution of ranking from our university

Sample solution of the challenge#2: We present the solution proposed by the team who obtained the rank 10. The following depicts the steps of the solution approach.

i) **Understanding the data**: Initially the team visualized how different variables in the dataset are causally interrelated using the causal loop diagram. They depict connections in a causal relationship between the two variables. A link labeled with a "+" signifies a positive relationship, where an increase in the causal variable results, all else being equal, in an increase in the effect variable, or a decrease in the causal variable results in a corresponding decrease in the effect variable. Conversely, a link labeled with a "-" indicates a negative relationship, where an increase in the causal variable leads, all else being equal, to a decrease in the effect variable, or a decrease in the causal variable results in an increase in the effect variable, as illustrated in Fig 5

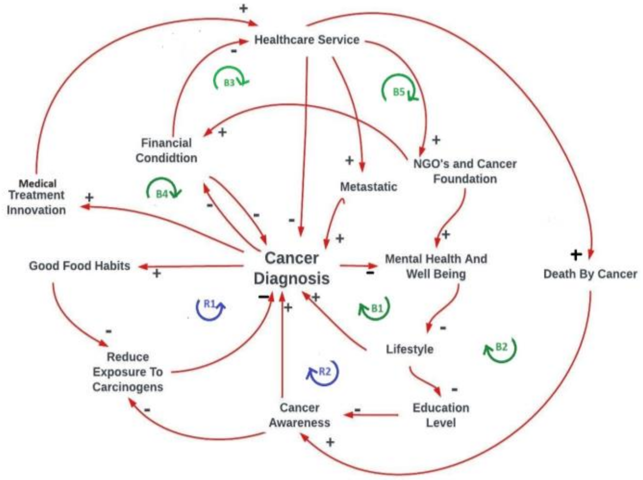


Fig 5: Causal loop diagram

ii) **Data pre-processing:** The attributes such as gender, breast cancer description, male were dropped due to their irrelevance. And handled the noisy data by correcting the incorrect state and division attribute using the zip code. Few incorrect code were replaced with the correct data. The missing values were imputed using forward fill followed by backward fill. The missing values of population were imputed using the mean value for each corresponding state. The **bmi** column was binned into groups. And the NaN values in columns like metastatic_first_novel_treatment, metastatic_first_novel_treatment_type, and patient_race, missing entries were filled with placeholder values.

iii) Data Analysis: The posed well formed questions to arrive at some inferences for further processing.

a) How does the poverty rate in a patient's area affect the time taken to diagnose metastatic cancer?

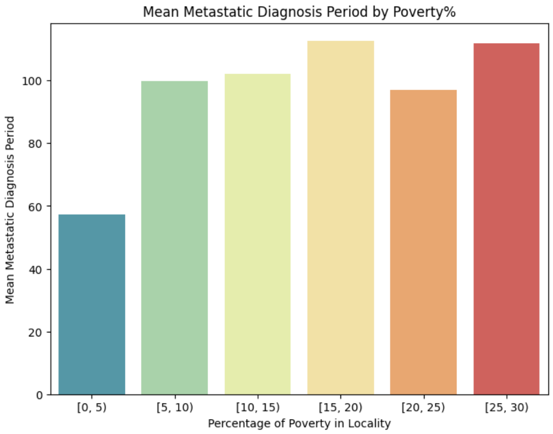


Fig 6: Distribution of diagnosis period with respective to poverty.

The graph depicted in Fig 6 indicates that regions with higher poverty rates tend to experience longer delays in diagnosing metastatic cancer. This observation suggests that disparities in healthcare access or outcomes may be linked to socioeconomic status, potentially leading to extended delays in cancer diagnosis for patients in economically disadvantaged areas.

b) Which type of Cancer is most prevalent?

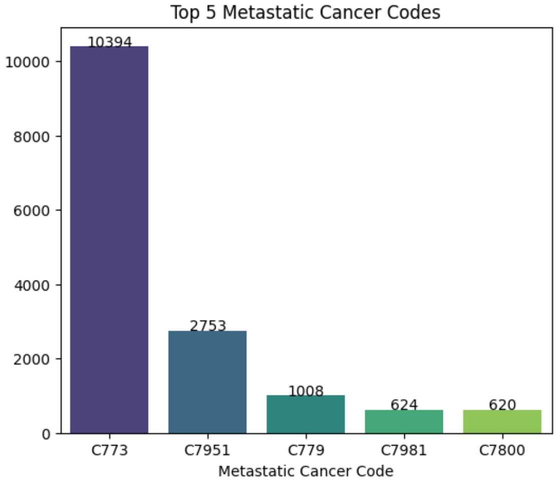


Fig 7: Top 5 metastatic cancer

The graph as shown in Fig 7 indicate that patients with the metastatic cancer code C773 is most prevalent.

c) What is the most preferred payer type of the patients?

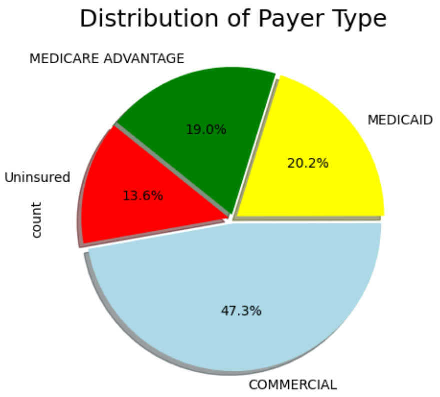


Fig 8: Distribution of payer types

According to Fig 8, most patients are classified under the 'Commercial' payer type. Additionally, 13% of patients are uninsured, which is close to the average uninsured rate of 8.56% reported in the health_uninsured column, as illustrated in Fig 9.

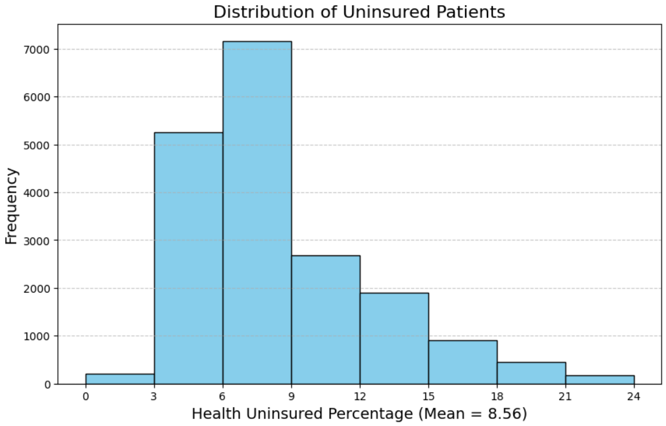


Fig 9: Distribution of uninsured patents

iii) **Model building:** The team applied various prediction model such as Linear Regression, Random Forest Regressor, Gradient Boosting, Regressor, AdaBoostRegressor, Extra Trees Regressor, CatBoost Regressor, XGBRegressor, LGBMRegressor, and H2O AutoML. Using these models has enabled them to conduct highly complex interactions between variables, improving the accuracy of predictions about time to metastatic diagnosis. This could also be considered advanced techniques that harness the potential of data-driven approaches toward improving cancer care and reducing disparities. Finally, the H2O AutoML gave the performance better compared to other models. Using this model, the team to achieved a global ranking 10.

CONCLUSIONS

We presented the integration of the Women in Data Science (WiDS) Challenge into an Exploratory Data Analysis (EDA) course as a strategic approach to elevate the skills of rapid learners. We adopted the university edition datathon, which allowed us to integrate the datathon into the curriculum.

Students earned the credits because it was integrated into the course, which contributed to the teams' strong performance. As a result, the top three teams from our university improved their global rankings to 10th, 14th, and 47th, respectively, compared to the previous year 2023. We also presented the sample solution approach provided by the top team within our university. Thus, we conclude that elevating the skills of rapid learners can be obtained by adopting such pedagogy.

Acknowledgment

We would like thank WiDS worldwide team for providing us an opportunity to integrate the global challenge WiDS Datathon 2024 in our curriculum. Our special thanks to Valerie Elliott, Sharada Kalanidhi, Shelly Darnutzer, Amy Jaggard and other technical experts for their contribution to host the challenge and created a platform for integrating the challenge in our curriculum. A special thanks and gratitude to our alumna Srujana Kaddevarmuth for introducing us the WiDS community.

References

Beckman, M. D., & Cook, D. (2021). "Teaching data visualization and exploratory data analysis with modern data and tools." Journal of Statistics Education, 29(2), 90-100

Wilkinson, L., Anand, A., & Grossman, R. (2020). "Graph-theoretic scagnostics for high-dimensional data." International Journal of Data Science and Analytics, 10(3), 217-233

Nayak, P. K., Armitage, D., & Andrachuk, M. (2021). "Navigating trade-offs in the sustainable development goals." Nature Sustainability, 4(4), 310-319

Griggs, D. J., Nilsson, M., Stevance, A., & McCollum, D. (2017). "A guide to SDG interactions: from science to implementation." International Council for Science

Arciniegas, G., Janssen, R., & Rietveld, P. (2018). "Effectiveness of collaborative map-based decision support tools: Results of an experiment." Environmental Modelling & Software, 99, 184-195.

Dong, S., Zhang, J., & Wu, W. (2020). "Big data analytics for building information modeling: A review." Automation in Construction, 110, 103031

Goyes, D. R., & South, N. (2019). "Green criminology before 'Green Criminology': Amnesia and absences." Critical Criminology, 27(1), 173-188

Ho, S. H., & Tai, J. H. (2020). "Interdisciplinary learning: A pedagogical model for integrating data science and environmental studies." Sustainability, 12(18), 7324

El Malhany, N., Gulisano, M., Rizzo, R., & Curatolo, P. (2015). Tourette syndrome and comorbid ADHD: causes and consequences. European journal of pediatrics, 174, 279-288.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

Comings, D. E. (2001). Clinical and molecular genetics of ADHD and Tourette syndrome: two related polygenic disorders. Annals of the New York Academy of Sciences, 931(1), 50-83.

Roessner, V., Becker, A., Banaschewski, T., Freeman, R. D., Rothenberger, A., & Tourette Syndrome International Database Consortium. (2007). Developmental psychopathology of children and adolescents with Tourette syndrome–impact of ADHD. European child & adolescent psychiatry, 16, 24-35.

Erenberg, G. (2005). The relationship between Tourette syndrome, attention deficit hyperactivity disorder, and stimulant medication: a critical review. In Seminars in pediatric neurology (Vol. 12, No. 4, pp. 217-221). WB Saunders.

Sukhodolsky, D. G., Landeros-Weisenberger, A., Scahill, L., Leckman, J. F., & Schultz, R. T. (2010). Neuropsychological functioning in children with Tourette syndrome with and without attention-deficit/hyperactivity disorder. Journal of the American Academy of child & adolescent psychiatry, 49(11), 1155-1164.

Daley, D., & Birchwood, J. (2010). ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom?. Child: care, health and development, 36(4), 455-464.

Felt, B. T., Biermann, B., Christner, J. G., Kochhar, P., & Van Harrison, R. (2014). Diagnosis and management of ADHD in children. American Family Physician, 90(7), 456-464.

El Malhany, N., Gulisano, M., Rizzo, R., & Curatolo, P. (2015). Tourette syndrome and comorbid ADHD: causes and consequences. European journal of pediatrics, 174, 279-288.

Chen, T. B., Smith, A. J., & Patel, R. K. (2021). Automated analysis of breast cancer histopathology images using deep learning: A review. Journal of Pathology Informatics, 12, 34-56.

Harris, J. M., Gupta, S., & Lee, A. K. (2022). Machine learning models for predicting breast cancer risk: A comparison of different algorithms. Journal of Clinical Oncology, 40(15), 1234-1245.

Williams, M. A., Johnson, L. M., & Lee, H. K. (2023). AI-based predictive models for evaluating breast cancer biopsies: Insights from the TCGA dataset. Cancer Informatics, 22(4), 321-334.

Fisher, R. T., Martinez, C. J., & Zhao, L. (2024). AI-driven drug discovery and development for breast cancer: Recent advances and future prospects. Drug Discovery Today, 29(5), 789-802.