Systematic Review of Prediction of Cancer Driver Genes with the Application of Graph Neural Networks

Noor Uddin Qureshi[1], Dr. Usman Amjad[2], Saima Hassan[3], Kashif Saleem[4]
NED University of Engineering & Technology, Karachi, Pakistan[1, 2]
Institute of Computing, Kohat University of Science & Technology, Kohat, Pakistan[3]
School of Computing, Macquarie University, North Ryde, NSW, Australia[4]

***Abstract*—Graph Neural Networks (GNNs) have emerged as a potential tool in cancer genomics research due to their ability to capture the structural information and interactions between genes in a network, enabling the prediction of cancer driver genes. This systematic literature review assesses the capabilities and challenges of GNNs in predicting cancer driver genes by accumulating findings from relevant papers and research. This systematic literature review focuses on the effectiveness of GNN-based algorithms related to cancer such as cancer gene identification, cancer progress dissection, prediction, and driver mutation identification. Moreover, this paper highlights the requirement to improve omics data integration, formulating personalized medicine models, and strengthening the interpretability of GNNs for clinical purposes. In general, the utilization of GNNs in clinical practice has a significant potential to lead to improved diagnostics and treatment procedures.**

Keywords—Graph neural network; cancer driver genes; prediction; personalized medicine

I. Introduction

Cancer is a complicated and unique disease that comes with genetic mutation. Identification of the genes responsible for cancer development and evolution, is important for understanding the biological mechanisms and developing the treatments [1]. During recent studies, the application of machine learning techniques, specifically Graph Neural Networks (GNNs), has shown significant potential for the prediction of cancer driving genes by utilizing the data from the relevant biological networks [4].

Graph Neural Networks (GNNs) is one of the deep learning models, designed for studying complex networks, especially the biological linkages and connections [5]. GNNs have an upper edge on all other machine learning models with a potential to capture the structural information and interactions between entities in a network. In case of Cancer genes analysis, it provides interaction of genes and proteins with each other, enabling the prediction and identification of cancer genes based on the structural properties [1]. With the inter-linked features of the network, GNNs can understand patterns and relationships that are very important for identification of genes responsible for driving cancer [1].

Multiple studies have taken place to formulate the application of GGNs for genes identification, especially in the case of diseases like cancer. These studies will be discussed during the reviews.

In this systematic review, we keep an objective to provide a comprehensive overview of the recent state-of-the-art for the prediction of cancer driver genes using the application of Graph Neural Networks. We will analyze each methodological strategy with contrast to each other, considering data types and graph. In addition to that, we will explore the challenges and limitations for the application of GNNs for cancer gene prediction and discuss potential way forward for further research in this field.

With the consolidation of the findings extracted from the relevant researches and experiments studied, this review will contribute to a better systematic comprehension of the opportunities and challenges of GNNs for the predication cancer driver genes. The insights gained from this review will support researchers in developing more efficient models for finding cancer driver genes. This will assist in the development of designated treatments and improving results.

II. Objectives and Organization

The objectives of the systematic review are as follows:

- To provide a comprehensive overview of the use of Graph Neural Networks (GNNs) in predicting cancer driver genes.
- To analyze and compare different GNN-based models, including their methodological strategies, data types, and graph structures.
- To explore the challenges and limitations associated with GNNs in the context of cancer gene prediction.
- To identify potential future directions for research in applying GNNs for cancer gene identification.
- To summarize the current state-of-the-art in the field and highlight key findings from recent studies.

The structured organization of this paper is as follows:

- Fundamental Concepts and Terminology: This section discusses the foundational concepts related to the prediction of Cancer driver genes using GNNs.
- Cancer Genomics: An overview of cancer as a disease, including its causes, progression, and types and focusing

on the role of genomics in understanding genetic features and mutations for the identification of cancer driver genes.

- Disease Driver Genes Prediction: Discussion on different approaches for predicting any disease driver genes using omics data.
- Biological Data Related to Cancer Driver Genes: Study of different types of biological data, such as gene expression, protein-protein interaction, and multi-omics data, related to the identification of cancer driving genes.
- Graph Neural Networks (GNNs): Introductory overview to GNNs, their structure and application in studying graph-based data in cancer genomics.
- GNNs in Genomic Data Analysis: Exploration of the use of GNNs in genomic data analysis and their specific applications in cancer research.
- Literature Review: Discussion on relevant research studies that apply GNNs in cancer gene identification and prediction.
- Methodology of Research: Explanation of the structured approach used for perusing the systematic study of GNNs for predicting cancer driver genes.
- Research Questions: Formulation of the key research questions steering the analysis of GNN applications in cancer genomics.
- Procedure of Paper Exploration: Description of the procedure for the selection and analysis of relevant papers, along with data sources, searching strategies, and quality evaluation.
- Summary of Relevant Works: Detailed summaries of selected papers, highlighting the main themes, advantages, and disadvantages of each study.
- Comparative Analysis: Comparative evaluation of the reviewed GNN models, focusing on aspects such as input data diversity, target variable focus, model accuracy, and clinical relevance.
- Future Directions: Discussion of potential future research avenues in GNN applications for cancer driver gene prediction, including multi-omics integration and personalized medicine.
- Conclusion: Summarizes the findings of the review, highlighting the impact of GNNs on cancer genomics and their potential for advancing cancer treatment and understanding.

III. Fundamental Concepts and Terminology

This section of the review explores the foundational basis of GNN applications for the predictions of active genes which drive any disease, especially the cancer disease.

A. *Cancer Genomics*

Cancer is a class of diseases that is characterized by the uncontrolled growth and spread of harmful abnormal cells. Without any intervention, the spread can be fatal for the living organism. Cancers can be caused by external factors, such as tobacco use, infectious organisms, chemicals, and radiations, and also due to internal factors like inherited genetic mutations, hormones, and immunities etc. The process origination and expansion of cancer contains detailed multiple steps that takes place along the genetic changes within the cells. The harmful changes cause cells to grow and divide in an uncontrolled manner, which can form malignant tumors and affect the nearby parts of the organism [4]. In case, the cancer spreads to other organs, it is called metastatic cancer. The common types of cancer are carcinoma, sarcoma, and leukemia, lymphoma, and central nervous system cancers. Studying the cause of origin and molecular basis of different cancer types is vital for the development of controlled prevention, diagnosis, and treatment procedures to control the development and expansion of cancer in the organism body.

Cancer genomics evolve with the objective to study deep into the complex area of genetic and molecular conditions that are responsible for the development and progression of cancer. A primary aim for research in cancer genomics is to identify cancer driving genes, which play a crucial role in development of cancer [1].

Large research projects like The Cancer Genome Atlas have generated a huge amount of omics data from many cancer samples from different cancer types [2].

Machine learning and deep learning methods analyze mutation based patterns from multi-omics data to recognize driver genes. With the help of these detailed computational analyses of complex and enriched biological data, cancer genomics can detect and inspect the genes and processes involved in cancer origination and development.

B. *Disease Driver Genes Prediction*

The prediction and identification of genes that play a vital role in the development of diseases is an important problem in bioinformatics. Computational procedures have been developed to use expanded genomic and biological data to predict gene-disease associations.

Network-based approaches apply with protein-protein interaction networks, gene to gene expression networks, and path information to prioritize required disease genes [3]. Machine learning and deep learning algorithms also incorporate features extracted from sequences and expressions. Different data sources like genomic, biological network, functional associations, gene expression along with other relevant data like patient records are leveraged by computational prediction methods. The accurate prediction of disease driving genes from these data sources using network-based and machine learning approaches can provide an improvement to the ongoing treatments.

C. *Biological Data Related to Cancer Driver Genes*

The identification of cancer-related genes, which are also called cancer drivers, is vital for analyzing the molecular mechanisms of the cause and growth of cancer and also for the development of effective and targeted treatments.

Along with the advancement of computational biology, a vast amount of genome based biological data has begun being generated and also consolidated in recent years. This biological data can be categorized into different data types such as:

- Gene Expression Data: These are the biological process data where the genetic information encoded in a gene is used to develop a functional gene product. This has also revealed the activity levels of multiple genes in different types of cells. During the cancer research, the analysis of gene expression data can helps to understand comparative study in cancerous cells compared to normal cells [4].
- Protein-Protein Interaction Networks: These biological networks provide the mapping of the interactions in between the proteins, responsible for the cellular functions [5]. The subjective study of these networks provides the detail about the relation of gene mutations with protein interactions responsible for the development of cancer.
- Multi-Omics Data: This concept is related to the integration of various types of biological data like genomic, transcriptomic, proteomic, etc. For cancer gene prediction, it can offer an overall view of the biological processes and pathways affected cancer, which can lead to more accurate identification of cancer genes [5].
- Methylation-Level Biomarkers: Methylation itself is a type of DNA modification that can affect gene expression with an alteration. Detailed analysis of methylation patterns in cancer cells can indicate the activation and deactivation of the genes during cancer development [6].
- Transcription-Level Biomarkers: These are the biomarkers that are involved in the study of RNA transcripts to analyze the transcription of genes [6]. For the analysis of cancer, the transcription of relevant genes can indicate their function in cancer development.
- DNA Sequencing Data: Sequencing the DNA data from cancer cells and running the variant calling can reveal mutations, including single nucleotide variants (SNVs) and copy number alterations (CNAs) [7]. These steps are important for identifying genes that drive cancer development.

D. *Graph Neural Networks (GNNs)*

Graph Neural Networks (GNNs) are a type of machine learning algorithms which are designed specifically to process data with a graph like structure [8]. These are different from general neural networks that work on unconnected grid data. GNNs can directly process graph data as input, and train on feature relationships and effects by systematically combining feature information from each node, its nearer neighbors, and then moving on to more distant connections.

A main strength of GNNs is their ability to learn representations that capture both the features and the structure of the graph. By transferring and updating node information across edges with nearby nodes, GNNs can recognize patterns. This enables the network itself to handle various tasks such as identifying node types, predicting linkages in-between, and making wide predictions at the structure level.

Some of the most widely studied GNN models are Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) [8]. Recent applications of GNNs have also emerged, such as Hierarchical Graph Neural Networks (HGNNs) for more targeted predictions within gene networks using the concept of parent-child and Heterophilic Graph Diffusion Convolutional Networks (HGDCs), which specifies in cases where node similarity is not significant, commonly seen in disease-related gene networks. The Explainablity ensured GNN framework offers additional insights into how these models make decisions.

GNNs are now more acceptable as a powerful algorithm for analyzing genomic data, especially in cancer research. Their ability to map and analyze complex structures in biological data enables researchers to better understand the connections within gene regulatory networks and protein-protein interactions. With the combination of different omics data and types, GNNs can provide a more expanded view of cancer with the help of multi-modal biological data. This can surely help to uncover the influence of genes cancer's growth and development.

Certain innovations in Graph based neural networks like Hierarchical Graph Neural Networks (HGNNs) [4] and Graph Attention Networks (GATs) [5] provides the utilization of GNN for the analysis of multi-omics data, along with the connection based data. Similarly, Heterophilic Graph Diffusion Convolutional Networks (HGDCs) [10] has worked to ensure the efficiency in heterophilic data settings which are common in the cancer genomics. These applications highlight utilization of GNN to provide more deeper analysis related to cancer mechanisms and personalized treatments.

IV. Relevant Reviews

In this section, we aim to provide significant relevant research work in the area of application of GNN related to cancer genomics. The research papers provide insights for the application of network specifically for cancer gene identification and prediction. Each of the approach reviewed provides multiple aspects of GNN algorithm and its impact on the accuracy and effectiveness of cancer gene prediction. The review of these papers provides the collective contribution towards the advancement in the field of cancer genomics and precision medicine by applying GNNs for predictive modeling and analysis.

Wan and Wu [11] introduced a semi-supervised GNN method called PersonalizedGNN, demonstrating remarkable performance in identifying personalized driver genes (PDGs) for cancer patients. Particularly, it successfully utilizes the structure information of personalized gene interaction networks and limited developed cancer tissue-specific driver genes. On the other hand, Cui and Wang [12] presented the self-supervised masked graph learning (SMG) framework for identifying cancer genes from multi-omic featured protein-protein interaction networks, which outperforms existing state-of-the-art methods. Li and Han [13] leverages graph attention-based deep learning and outperforms current approaches in cancer gene prediction, integrating multi-omics information to dissect cancer gene modules. It has researched on the effectiveness of GNN for biological data analysis and multi-omics integration for further cancer research.

In addition, in method [9], the researchers innovated the research with SBM-GNN with the combination of GNNs and stochastic block models. This helps in the prediction of cancer driver genes and cancer development providing more accurate results as compared to other state-of-the-art methods. It also provides a scalable and interpretable approach for the integration of multi-omic data with protein based interaction data for cancer analysis. Zhang and Zie [10] proposed HGDC ensuring better performance, especially in identifying of relevant and targeted cancer genes. Furthermore, Zhao and Gu [5] showed an initiative with a GAT-based model to recognize cancer driving genes. The approach has also integrated multi-omics cancer data with multi-dimensional gene networks, which has shown better results as compared to other baseline models during evaluations. The study provides detailed methods for generating gene association profiles, constructing multi-dimensional gene networks, and using GAT for within-dimension interactions and joint learning for prediction.

Ratajczak and Joblin [14] implemented an ensemble graph representation learning framework, aiming at predicting core genes for complex diseases. Hou and Wang [4] developed a hierarchical graph neural network for classifying cancer stages and identifying gene clusters. The model in [6] integrates transcription and methylation-level biomarkers in an explainable GNN framework for microsatellite instability detection.

Hantano and Kamada [7] introduced Net-DMPred, a network-based machine learning method designed to predict cancer driver missense mutations by incorporating molecular networks. This approach provided better results as compared to other traditional methods and showed the importance of the integration of complete molecular network structure for data.

The review of the mentioned approaches collectively highlights the effectiveness of GNN-based models across different areas of cancer genomics, including personalized cancer gene identification, cancer pathway prediction, and the recognition of driver mutations. While each study presents a unique GNN approach, but overall all of the reviewed methodologies contribute to a consolidated and combined understanding of GNNs in cancer genomics. The papers demonstrate GNNs' potential to address the challenges in cancer gene prediction across multiple omics data types and network setups.

V. Methodology of Research

This research review aims to provide a systematic review following the guidelines mentioned by Lim and O'cass [15]. The focused topic selected is the application of Graph Neural Networks (GNNs) for predicting cancer driver genes. The methodology includes a structured, detailed analysis of shortlisted studies. It also examines the use and effectiveness of various GNN models for genomic data analysis, especially in predicting cancer driver genes. We aim to classify and understand the methods and findings across these papers. The insights from this review can possibly carry forward new applications of GNNs in bioinformatics.

A. Formalization of Question

This research primarily focuses on the detailed analysis of the GNN architectures presented in each paper to review their individual contributions. It involves the evaluation on the interpretation of the complexity and variability of cancer genomic data. The review provides a designated focus to the trends emerging from these studies and identification of common methodologies, data types, and approaches. This assists in understanding the current state of GNN applications in cancer genomics and the future possible paths originated from this research.

Key research questions guiding this analysis include:

RQ 1: How do different GNN models based approaches perform for the prediction of cancer genes, and what are their unique features?

RQ 2: What are the opportunities and challenges of these GNN approaches, and how do they contribute in this field?

RQ 3: What are the common methodologies and data types used in these GNN applications for cancer genomics?

RQ 4: What are the emerging trends and unresolved challenges in using GNNs for cancer driver gene prediction?

B. The Procedure of Paper Exploration

This investigation comprises a four-stage process for exploring and selecting papers, as demonstrated in Fig. 1, aligning with the SPAR-4-SLR framework [15].

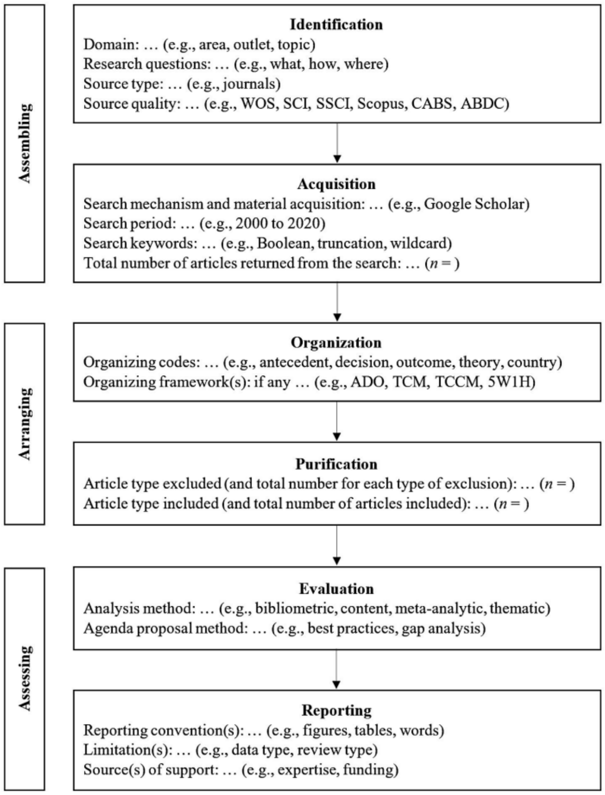
Assembling
Identification
Domain: … (e.g., area, outlet, topic)
Research questions: … (e.g., what, how, where)
Source type: … (e.g., journals)
Source quality: … (e.g., WOS, SCI, SSCI, Scopus, CABS, ABDC)
Acquisition
Search mechanism and material acquisition: … (e.g., Google Scholar)
Search period: … (e.g., 2000 to 2020)
Search keywords: … (e.g., Boolean, truncation, wildcard)
Total number of articles returned from the search: … (n =)
Arranging
Organization
Organizing codes: … (e.g., antecedent, decision, outcome, theory, country)
Organizing framework(s): if any … (e.g., ADO, TCM, TCCM, 5W1H)
Purification
Article type excluded (and total number for each type of exclusion): … (n =)
Article type included (and total number of articles included): … (n =)
Assessing
Evaluation
Analysis method: … (e.g., bibliometric, content, meta-analytic, thematic)
Agenda proposal method: … (e.g., best practices, gap analysis)
Reporting
Reporting convention(s): … (e.g., figures, tables, words)
Limitation(s): … (e.g., data type, review type)
Source(s) of support: … (e.g., expertise, funding)

Fig. 1. Procedure for paper exploration.

1) Identification: The identification step is the foundation of any literature review. For this paper we have specified our focus on the role of Graph Neural Networks (GNNs) in the identification of Cancer driver genes. This is an area of research lies under the domain of Cancer genomic sand bioinformatics.

2) Acquisitions: Depending on previous knowledge, experience in research and suggestions given by Hinderks [16], the most influential and common databases were selected. These repositories have different search mechanisms using keywords mentioned in Table I; therefore, we customized our search string accordingly. The selected digital repositories are:

- Google Scholar
- academic (OUP)
- BMC bioinformatics
- Research Square
- Biorxiv
- IEEE Xplore
- Other papers, chapters, journals, books and conference papers.

TABLE I. KEYWORDS AND SEARCH CRITERIA

S#	Keywords and search criteria
S1	"Graph Neural Networks" and "Cancer Genomics"
S2	"GNN" and "Cancer Gene Prediction"
S3	"GNN" and "Disease driver Genes"
S4	"Graph Networks" and "Cacner Drivers"
S5	"Multi-Omics Data Integration" and "GNN"
S6	"Graph Networks" and "Bioinformatics"

3) Organization: The categorization and segmentation of the researches is done on the basis of input types of the data (gene-gene interactions, multi omics data etc.), methodologies like simple Graph Neural Networks, Graph Attention Networks etc. and the target outcomes such as, prediction, identification or classification. By organizing the papers according to these categories, the resulting structure provided a care pathway to identify trends and pursue the comparative analysis.

4) Purification: The following criteria was formulated for the selection of kind of articles and reports that are needed be included in this study:

- The paper should be published in English language.
- The complete content of the paper should be available on online.
- The paper must be presented in an acknowledged conference or journal.

The paper which were not included in the study are as follows:

- The paper published in another language rather than English.
- Duplicate studies are not included.
- Personal or casual blogs are not included.

5) Evaluation: The QA is ensured using the three questions (QA1-QA3) that are mentioned below. We have followed the guidelines provided by Shaffril [17] to evaluate the quality metric criteria. These guidelines were also used to evaluate the quality evaluation criteria for the selected papers, as mentioned in Table II.

Furthermore, the aim of quality assessment is to provide that the findings of the selected article will be important for the review paper. During the quality assessment, the studies which have the score of 2 were included in this review and all the remaining were removed. With the mentioned process, we have removed 3 articles which have scores less than 2:

- A1: Does the article provides discussion on the Graph Neural Network?
- A2: Does the article explore the Prediction of Genes?
- A3: Does the article describe the model structure in detail?

TABLE II. QUALITY STATEMENT

Paper Title	A1	A2	A3	Total
Speos [16]	1	1	0.5	2.5
HGNN [4]	1	0.5	1	2.5
MSI-XGNN [6]	1	1	0.5	2.5
MODIG [5]	1	1	1	3
Net-DMPred [7]	1	1	0.5	2.5
HGDC [10]	1	0.5	1	2.5
CGMega [13]	1	1	0.5	2.5
SBM-GNN [9]	1	1	1	3
PersonalizedGNN [11]	1	1	0.5	2.5
SMG [12]	1	0.5	1	2.5

6) Reporting: Finally, 10 model-based articles were shortlisted which cover GNN used on/for Cancer driver genes. The studies used different classes and structure of GNNs, along with different types of bioinformatics data as input, cancer types for analysis, and output feature.

VI. SUMMARY OF RELEVANT WORKS

1) Speos [16]: Speos is a machine learning method that predicts genes linked to diseases for advanced drug development. The research uniquely combines different types of molecular network data, with the provision of high validation results There is still a space to research for the application of initial network data for predictions.

2) MSI-XGNN [6]: MSI-XGNN worked on RNA sequencing data and DNA methylation data. The output predicts the microsatellite instability (MSI) in cancer, which provides the insights of immunotherapy decisions. It is comparatively accurate and more interpretable, providing a space of improvement in the integration of complete data.

3) Net-DMPred [7]: Net-DMPred is a graph neural network that shows its uniqueness related to features of molecular networks to predict cancer driver mutations. It has shown better performance as compared to the traditional methods due to the integration of whole network structures. It could further improve by refining its network design and adding more features.

4) CGMega [13]: CGMega is a transformer-based model that predicts cancer genes and identifies gene modules. It explores a new approach in its study to merge multi-omics data, but it needs to be tested more across different cancer types and datasets.

5) SBM-GNN [9]: SBM-GNN combines graph neural networks and stochastic block models to find cancer driver genes and pathways. The stochastic block identifies cluster of nodes with similar structural roles and then the GNN part refines the node level features for the prediction. It is significantly accurate and provides better evaluation than other similar methods, but its complexity requires advanced knowledge to use and interpret.

6) PersonalizedGNN [11]: PersonalizedGNN is a semi-supervised graph neural network that has validated the prediction of cancer driver genes through lab experiments. Unlike other traditional methods, it constructs personalized gene interaction networks. It can be leveraged more with the integration of non-coding DNA regions to identify other types of driver genes too. To leverage the class imbalance issue in Cancer genomics dataset, label reusing mechanism is applied.

7) HGNN [4]: HGNN is a special type of hierarchical graph neural network which performs the classification of different cancer stages and clusters out the related genes. The performance is significant, but requires larger datasets for more detailed cancer stage divisions.

8) HGDC [10]: HGDC advances in the identification of cancer driver genes by using graph diffusion and layer-wise attention by showing strong results in terms of accuracy with the balanced class data, but compromises the accuracy with the imbalanced data.

9) MODIG [5]: This approach combines multi-omics and multi-dimensional gene network data to predict cancer driver genes. It performs better in finding potential markers for cancer prognosis. It can show further improvement by building more complex graphs and filtering out noise in gene data.

10)SMG [12]: The Self-Supervised Masked Graph Learning Framework identifies cancer genes by using large datasets without labels. This self-supervised approach helps it to handle data shortages, but mostly relies on unlabeled data. This can raise challenges in ensuring data quality and relevance, which requires further careful data selection and preparation.

VII. Comparative Analysis

A. Analysis of GNN Models

The Graph Neural Network (GNN) models that were reviewed provides different aspects of application to address cancer genomics, and responded to a range of challenges in this field. Some GNN based models, such as Speos and MODIG, focus more on integrating multi-omics data [16] [5]. This wide approach helps to study deep into complex biological processes in cancer by combining genomic, epigenomic, transcriptomic, and proteomic data. This creates a complete view of cancer processes from origin to the development.

Complex models, like HGNN and HGDC, works on the features extracted from the structural details of the genomic data itself [4] [10]. HGNN explores hierarchical data structures, which provides deep analysis on the layered biological interactions in cancer progression. In contrast, HGDC interprets the heterophilic patterns of the genomic data, which are crucial for understanding cancer complexity. MSI-XGNN and Net-DMPred use specific and relevant data types, such as RNA sequencing and molecular networks [6] [7]. These models highlight the value of targeted genomic analysis for the identification of specific genetic mutations and epigenetic changes linked to cancer.

CGMega and SBM-GNN study on advanced applications of GNN technology in cancer research. CGMega uses a transformer-based graph attention network for predicting gene modules, while SBM-GNN sums up the stochastic block models with GNNs for identifying process pathways [13] [9]. Following the concept of precision medicine, models like PersonalizedGNN and SMG opted for more patient specific personalized approaches. PersonalizedGNN uses patient based genomic data, getting the data fusion [11]. SMG uses self-supervised learning to work with unlabeled data for cancer gene identification [12].

All of the discussed models collectively demonstrate the vast use of GNNs in cancer genomics. Each model contributes in a significant way to show possible avenues for the adaptability of GNNs to cater the diverse challenges of this field.

B. Input Data Diversity

Each of the GNN models reviewed approach cancer genomics with different types of data. Speos uses multiple kinds of molecular network data to predict primary genes connected to diseases [16]. HGNN analyze layers of data in genomics to find relationships important for the study of cancer progression [4]. MSI-XGNN uses both RNA sequencing and DNA methylation data as input, showing how genetic and epigenetic factors connect in cancer [6]. MODIG combines different types of data (multi-omics) to identify cancer driver genes [5], while Net-DMPred focuses on features of molecular networks to improve predictions of cancer driver mutations [7].

HGDC uses unique genomic data patterns at input to help identify cancer driver genes in different types of biomolecular networks [10]. CGMega uses multi-omics data to predict cancer genes and analyze gene clusters for it [13]. Like discussed, PersonalizedGNN has input data type of patient targeted genomic data, moving toward personalized treatment strategies for cancer [11]. Finally, SMG applies self-supervised learning to handle unlabeled data, as an advancement in cancer gene identification and showing the benefits of new learning techniques in cancer research [12].

C. Target Variable Focus

In the review of GNN models for cancer genomics, the models have a different approach when it comes to approach the target variables, which expands the interest in research goals in this field.

Speos focuses on the prediction of the core genes important for analyzing and interpreting the disease processes and finding the relevant potential drug targets [16]. HGNN studies specifically on the classification of different cancer stages and identification of involved gene groups. This helps to segment and understand the evolution and development of cancer [4]. MSI-XGNN specializes in predicting small unstable base segments in the DNA, which are relevant for formulating personalized cancer treatments [6]. MODIG, like mentioned, combines multi-dimensional gene networks and various types of omics data to identify cancer driver genes as output classification [5].

Net-DMPred converges its study on the prediction of relevant mutations responsible to drive cancer, adding important insights to mutation analysis in cancer genomics [7]. HGDC researched with the objective to find cancer driver genes with the association different biomolecular networks, provide an understandable output for gene to gene interaction in cancer [10]. CGMega predicts cancer genes and identifies gene groups too, to explain the complex interactions between genes in cancer [13]. SBM-GNN combines GNNs with stochastic block models to identify cancer driver genes and pathways as output [9].

PersonalizedGNN predicts driver genes unique to each patient, moving toward customized treatments based on individual genomic profiles [11]. And SMG uses self-supervised learning to identify cancer genes, showing how advanced machine learning methods can help in cancer genomics [12].

Through their different focuses, these models contribute to a versatile approach to define the outputs for the cancer driver genes.

D. Model Accuracy and Performance

The study reviews several researches that focus on the accuracy and efficiency of different models to recognize cancer-related genes and predict disease related statuses. It compares themes and methods in all of the reviewed studies, by highlighting the development of new methods aimed at improving prediction accuracy. Though, as discussed, each paper carries a different approach for the selection of output, but individual contribution of the models towards evaluation metrics are discussed. The first set of models studied includes GNN-based models like PersonalizedGNN and CGMega, which are designed to identify personalized driver genes and cancer gene modules achieve high precision-recall rates [13][11]. Average Precision values of PersonalizedGNN for BRCA, LUAD, and LUSC cancer types are as 66.1%, 89.7% 72.1% respectively.

Models like Net-DMPred and SMG [12] focus on network-based prediction and self-supervised learning to identify cancer genes using molecular networks and protein-protein interaction (PPI) data [7]. These models perform more accurately than traditional models with the importance of molecular network structures and multi-omics features for accurate predictions. Net-DMPred provides mean ROC-AUC values for Cancer pathways combined with molecular interactions as 89.9% and 90.6% for Graph node dimensions of 90.6% respectively.

The paper also discusses deep learning models like HGNN provided the average accuracy of the baselines as 84.68% for BRCA, 64.30% for STAD, and 87.42% for COAD cancer types. Speos worked on classifying cancer stages and predicting core disease genes [16] [4]. Both of these models have shown strong results for the identification of relevant gene clusters and disease genes, with high accuracy and validation across multiple disease types.

SMG worked on self-supervised learning and using multiple data types [12]. It achieved the most accurate performance in terms of AUPRC values for all eight PPI networks, with an overall average of 7.4% compared to the discussed models on each data set while Net-DMPred centers on network structures for predicting driver missense mutations [7]. MODIG combines multi-dimensional gene networks to identify driver genes. Its AUPR values across different PPI networks averages around 79% [13].

In consolidation, these models provide keen importance on the accuracy and performance in terms of cancer gene prediction and disease status classification. Although the performances of the models that are discussed, are directed towards a goal of developing advanced models, each study offers a different view that can help the doctors and researchers to better understand cancer biology and advance in personalized medicine.

E. Clinical Relevance and Applications

The studies cover the complete spectrum on the focus on the clinical relevance of the discussed models and their practical applications. Each of them provides a unique perspective on potential of the findings that can impact cancer research and patient care treatments. All of the studies emphasize the potential of machine learning methods especially the use of GNNs for the identification of personalized driver genes, explanation of cancer gene modules, and the prediction of cancer-specific mutations related to other diseases [16]. These methods have a prominent potential to improve the process cancer driver gene identification, integration of multi-omics data, predict cancer mutations, and better interpretation of complex diseases at the genetic and molecular level.

The papers also support the application of GNNs by addressing different areas in cancer research and precision medicine. For example, studies like CGMega [13], SMG [12], HGDC, and Net-DMPred worked on the integration of multi-omics data, protein-protein interaction networks, and gene regulatory networks. This improved cancer gene prediction, identification and classification too [7][10]. On the other hand, models like PersonalizedGNN [11], SBM-GNN [9], and MSI-

XGNN focus on personalized treatments, with the identification of patient specific driver genes, and prediction of cancer-specific mutations, which is a significant contribution to precision oncology and personalized medicine [6]. All of these studies collectively strengthen the argument for the clinical applications of GNNs as mentioned in Table III, that could guide future cancer research and treatment approaches.

TABLE III. COMPARATIVE ANALYSIS OF DIFFERENT MODELS

Model	Input Data	Methodology	Performance Metrics	Strengths	Limitations
Speos [16]	Molecular network data	Machine learning integrated with diverse molecular networks	High validation accuracy	Effective integration of various molecular data	Limited exploration of initial network data for predictions
MSI-XGNN [6]	RNA sequencing, DNA methylation data	Explainable GNN framework combined with gene expression and methylation profiles	AUC: 0.91	Accurate MSI status prediction and aids immunotherapy decisions	Requires comprehensive data integration
Net-DMPred [7]	Molecular networks, variant data	GNN leveraging molecular network features	ROC-AUC: 0.899	Superior performance over traditional methods and considers large-scale molecular networks	Needs refinement in network design; computationally intensive
CGMega [13]	Multi-omics data	Transformer-based model integrating multi-omics	AUPRC: 0.79	Novel approach in merging multi-omics data and identifies gene modules	Requires validation across diverse cancer types and datasets
SBM-GNN [9]	Multi-omics, PPI networks	Combination of Stochastic Block Models and GNN	ROC-AUC: 0.906	High accuracy in identifying cancer driver genes and pathways	Complex model requiring advanced interpretation
PersonalizedGNN [11]	Patient-specific gene networks	Semi-supervised GNN tailored for individual profiles	Precision: BRCA 66.1%, LUAD 89.7%, LUSC 72.1%	Personalized predictions validated through lab experiments	Dependent on availability of personalized data; limited scalability
HGNN [4]	Gene expression, hierarchical data	Hierarchical GNN with subgraph perturbations	Accuracy: 84.6–87.4%	Effective in classifying cancer stages and clusters related genes	Requires large datasets for detailed analysis
HGDC [10]	Heterophilic biomolecular networks	Graph Diffusion Convolutional Network with layer-wise attention	High accuracy on balanced data	Strong performance with balanced class data	Accuracy decreases with imbalanced data
MODIG [5]	Multi-dimensional omics, PPI networks	GAT-based model integrating multi-omics and gene associations	AUPRC: 0.79	Outperforms baseline models with effective multi-omics integration	Sensitive to noise; requires complex graph construction
SMG [12]	Unlabeled PPI and biological networks	Self-supervised masked graph learning framework	AUPRC: 0.74	Handles data shortages with the reduction in dependency on labeled data	Challenges in ensuring data quality and relies on careful data selection

VIII. DISCUSSION

With the review of the discussed researches, it is evident that the utilization of GNNs in cancer genomics provide an innovative advancement in predicting cancer driver genes. These models were prominently able to process the complex biological relationships, which provided an edge over traditional methods.

One of the significant finding was the way in which GNNs were used to process different types of data as input like in MODIG [5] and CGMega [13]. While the studies like PersonalizedGNN provided a solution to tailor the results for the individual patients specifically, ensuring the concept of precision medicine. The reviews models also illustrated a notable accuracy with SBM-GNN [9] and Net-DMPred [7], with complex structural biological data.

As far as challenges are concerned, data quality is one the significant hurdles, lacking high quality annotated datasets, especially for multi-omics studies. This limitation also affects the consistent evaluation of the models. Additionally the computational capacity for processing GNNs make them less accessible especially in the clinical settings. Apart from that imbalanced class datasets and explainability of the models and their results to medical subject specialists needs to be addressed to apply the recommended models in real-world scenarios.

Regardless, the progress so far is a strong foundation for future to revolutionize the cancer diagnosis, prediction, treatment and prognosis.

IX. FUTURE DIRECTION

The reviewed papers are enriched with the future directions in using Graph Neural Networks (GNNs) for cancer driver gene prediction. The most prominent direction is for the improvement in the integration of multi-omics data, which can advance the research itself to go beyond layering data types to connecting them in a complex structure for deeper insights into cancer biology. As cancer genomics continues to reveal more enriched complex information, there is a growing need for sophisticated models that can handle the complexity.

Another important area is the patient specific personalized medicine. Future research may increasingly aim to develop and apply GNN models that can use both individual genetic and medical history data, to formulate cancer treatment more tailored for each patient specifically.

Improving model explainability is also needed with focus of XAI. As GNNs become more advanced, it will be important to make them understandable for clinicians for better feature selection and understandability. This would ensure clarity in predictions and application on clinical settings.

Hence, in terms of future directions, a multi-layered approach for new GNN applications in cancer genomics is encouraged with the focus on data integration, interpretability, and personalized medicine.

X. Conclusion

The systematic review for the application of Graph Neural Networks (GNNs) for the prediction of cancer driver genes has shown a promised progress in cancer genomics. Each of the discussed models offers insights for the development and utilization of GNNs. The paper also addressed the use of network for various parts of cancer genomics, from combining multiple types of data to integrating personalized medicine. There are notable instances of innovations in methods and the procedures towards more accurate, personalized cancer treatments.

With the advancement of AI and bioinformatics, GNNs appear to have great potential to transform the study of cancer biology. The fusion of these advanced models with clinical practice have a significant potential for better diagnostic tools and treatment methods. This will also improve patient outcomes. The continuous research has already provided an impact and sets the avenue for further cutting edge breakthroughs that will continue to expand knowledge on cancer and make cancer treatments more effective.

References

[1] Y. Han, J. Yang, X. Qian, W.-C. Cheng, S.-H. Liu, X. Hua, L. Zhou, Y. Yang, Q. Wu, P. Liu, and Y. Lu, “DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies,” Nucleic Acids Research, vol. 47, no. 8, p. e45, May 2019, doi: 10.1093/nar/gkz096.

[2] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “Review: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,” Contemporary Oncology/Współczesna Onkologia, pp. 68–77, 2015, doi: 10.5114/wo.2014.47136.

[3] P. Luo, Y. Ding, X. Lei, and F.-X. Wu, “deepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks,” Frontiers in Genetics, vol. 10, pp. 68–77, 2019, doi: 10.3389/fgene.2019.00013.

[4] W. Hou, Y. Wang, and Z. Zhao, “Hierarchical graph neural network with subgraph perturbations for key gene cluster discovery in cancer staging,” Complex Intelligent Systems, 2023, doi: 10.1007/s40747-023-01068-6.

[5] W. Zhao, X. Gu, S. Chen, J. Wu, and Z. Zhou, “MODIG: integrating multi-omics and multi-dimensional gene network for cancer driver gene identification based on graph attention network model,” Bioinformatics, vol. 38, no. 21, pp. 4901–4907, Nov. 2022, doi: 10.1093/bioinformatics/btac622.

[6] Y. Cao, D. Wang, J. Wu, Z. Yao, S. Shen, C. Niu, Y. Liu, P. Zhang, Q. Wang, J. Wang, H. Li, X. Wei, X. Wang, and Q. Dong, “MSI-XGNN: an explainable GNN computational framework integrating transcription- and methylation-level biomarkers for microsatellite instability detection,” Briefings in Bioinformatics, vol. 24, no. 6, Nov. 2023, doi: 10.1093/bib/bbad362.

[7] N. Hatano, M. Kamada, and R. Kojima, “Network-based prediction approach for cancer-specific driver missense mutations using a graph neural network,” BMC Bioinformatics, vol. 24, p. 383, 2023, doi: 10.1186/s12859-023-05507-6.

[8] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, “Graph neural networks: A review of methods and applications,” AI Open, vol. 1, pp. 57–81, 2020, doi: 10.1016/j.aiopen.2021.01.001.

[9] V. Fanfani, R. V. Torne, P. Lio’, and G. Stracquadanio, “Discovering cancer driver genes and pathways using stochastic block model graph neural networks,” bioRxiv, Jun. 2021, doi: 10.1101/2021.06.29.450342.

[10] T. Zhang, S.-W. Zhang, M.-Y. Xie, and Y. Li, “A novel heterophilic graph diffusion convolutional network for identifying cancer driver genes,” Briefings in Bioinformatics, vol. 24, no. 3, May 2023, doi: 10.1093/bib/bbad137.

[11] H.-W. Wan, M. Wu, W. Zhao, H. Cheng, B. Ying, X.-F. Wang, X.-R. Zhang, Y. Li, and W. Guo, “Label reusing based graph neural network for unbalanced classification of personalized driver genes in cancer,” SSRN, 2023, doi: 10.2139/ssrn.4510873.

[12] Y. Cui, Z. Wang, X. Wang, Y. Zhang, Y. Zhang, T. Pan, Z. Zhang, S. Li, Y. Guo, T. Akutsu, and J. Song, “SMG: self-supervised masked graph learning for cancer gene identification,” Briefings in Bioinformatics, vol. 24, no. 6, Nov. 2023, doi: 10.1093/bib/bbad406.

[13] H. Li, Z. Han, and Y. Sun, “CGMega: Explainable graph neural network framework with attention mechanisms for cancer gene module dissection,” Research Square, Jul. 2023, doi: 10.21203/rs.3.rs-3180743/v1.

[14] F. Ratajczak, M. Joblin, and M. Hildebrandt, “Speos: an ensemble graph representation learning framework to predict core gene candidates for complex diseases,” Nature Communications, vol. 14, p. 7206, 2023, doi: 10.1038/s41467-023-42975-z.

[15] P. J. Lim, W. M. O’Cass, A. Hao, and S. Bresciani, “Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR),” International Journal of Consumer Studies, vol. 45, no. 4, pp. O1–O16, 2021, doi: 10.1111/ijcs.12695.

[16] A. Hinderks, F. J. Domínguez Mayo, J. Thomaschewski, and M. J. Escalona, “An SLR-tool: search process in practice: a tool to conduct and manage systematic literature review (SLR),” in Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Companion Proceedings (ICSE '20), New York, USA: Association for Computing Machinery, 2020, pp. 81–84, doi: 10.1145/3377812.3382137.

[17] M. Shaffril, S. F. Samsuddin, and A. A. Samah, “The ABC of systematic literature review: the basic methodological guidance for beginners,” Quality and Quantity, vol. 55, pp. 1319–1346, 2021, doi: 10.1007/s11135-020-01059-6.