# Metastatic Site Prediction in Breast Cancer using Omics Knowledge Graph and Pattern Mining with Kirchhoff's Law Traversal

Alokkumar Jha[a,b,c,1,*], Yasar Khan[a,1], Ratnesh Sahay[a,1]

[a] *Data Science Institute, National University of Ireland, IDA Business park, Lower Dangan, Galway, Ireland, H91EWX9*
[b] *Cardiovascular Institute, Stanford University School of Medicine, Palo Alto, CA*
[c] *Stanford Cancer Institute, Stanford University School of Medicine, Palo Alto, CA*

**Abstract**

Prediction of metastatic sites from the primary site of origin is a impugn task in breast cancer (BRCA). Multi-dimensionality of such metastatic sites - bone, lung, kidney, and brain, using large-scale multi-dimensional Poly-Omics (Transcriptomics, Proteomics and Metabolomics) data of various type, for example, CNV (Copy number variation), GE (Gene expression), DNA methylation, pathways, and drugs with clinical associations makes classification of metastasis a multi-faceted challenge. In this paper, we have approached the above problem in three steps; 1) Applied Linked data and semantic web to build Poly-Omics data as knowledge graphs and termed them as cancer decision network; 2) Reduced the dimensionality of data using Graph Pattern Mining and explained gene rewiring in cancer decision network by first time using Kirchhoff's law for knowledge or any graph traversal ; 3) Established ruled based modeling to understand the essential -Omics data from poly-Omics for breast cancer progression 4) Predicted the disease's metastatic site using Kirchhoff's knowledge graphs as a hidden layer in the graph convolution neural network(GCNN). The features (genes) extracted by applying Kirchhoff's law on knowledge graphs are used to predict disease relapse site with 91.9% AUC (Area Under Curve) and performed detailed evaluation against the state-of-the-art approaches. The novelty of our

*Corresponding Author

approach is in the creation of RDF knowledge graphs from the poly-omics, such as the drug, disease, target(gene/protein), pathways and application of Kirchhoff's law on knowledge graph to and the first approach to predict metastatic site from the primary tumor. Further, we have applied the rule-based knowledge graph using graph convolution neural network for metastasis site prediction makes the even classification novel.
*Keywords:* Knowledge Graph, Cancer Genomics, Deep Learning, Semantic Web Linked data

---

## 1. Introduction

Metastatic breast cancer is one of the most lethal diseases concerning treatment and relapse status. Approximately 154,794 patients from stage 1-3 were expected to have *MBC(Metastatic breast cancer)*[1]. To develop a classifier to address this problem we need to integrate the heterogeneous multi-dimensional data. This data is usually genomics data which covers DNA methylation, copy number variation, gene expression, various disease, and drug databases. The biomarker discovery of the genomics era has been usually gene expression which implies numerical data driven. The number of biomarkers between two studies for the same cancer types provides two different lists of genes. Multi-class and multi-feature cancer type implies a biased feature selection if not treated on an individual data source. Further, in metastatic pathological subtypes behave differently concerning metastatic sites in BRCA which leads to **Insufficient data or redundant data: Knowledge graphs can be a solution**. Further, Overlapping biomarkers in the primary site and metastatic sites helps to understand visceral metastasis this implies **Knowledge Graph should be more than a simple graph embedded: Rule-based Modeling**). The multi-organ vs. primary site study facilitates to study the ER/PR/HER2(+/-) regimes. Such classification makes this problem multi-class classification and requires to adapt various classification model. Thus, the application of tensor using a convolution neural network will establish a prototype for many can-

cer types. In graph mining for systems biomedicine, the path of genes across the metastatic tissues and change in their expression and associated pathways provides metastatic biomarkers leads to (**feature selection should be dependent on graph pattern than a learning curve: Since Disease is mechanism driven**). Further, in case of multiple metastases within the same organ(hepatic metastasis) it is critical to know the relative expression change in hepatic and multi-site metastasis. The problem arises the **(requirement of time-dependent Knowledge graph where each node carries the previous status and changed status)**.

## 2. Background

In the heterogeneous datasets, such a genomics data where knowledge distribution and knowledge completion is a big issue, Knowledge graph is going to provide the first-order enriched knowledge. Here "first-order" means the enrichment of data based on genomic entities match such as gene symbol and probe Ids in methylation data. However, in multi-predictive models, such a metastatic breast cancer knowledge graph provides first-order knowledge which is a combination of Omics data(genomics, drug, and disease). The diversity and subtyping of disease demanded to keep knowledge graph into various layers thus cataloged in subcategories such as tissue, drug, enzymes, mutations and drug effects, etc. Further, the importance of each layer of the knowledge graph may change based on cancer type or cancer subtype since each cancer type could be driven by one or more than one layers of knowledge graph. Further, cataloging and facilitating personalized model to each patient based on their treatment requires to have a knowledge graph regarding flexibility in the ranking of events. For example, in Triple Negative breast cancer mutations play a key role and will impact the performance of model more wherein other subtypes CNV variants need to have property. Similarly, the patient in treatment requires a drug response to be kept in priority. So Current static knowledge graphs are not serving the purpose. To achieve such a dynamic and multi-model knowledge graph, it requires a traversal

and rule-based modeling. Since rule derived by the domain can help to eliminate the noisy data while training the model for prediction and traversal based on rules requires to have a model which respects the weight of each and effect on overall knowledge graph. Doing this will help to identify the effectiveness of the knowledge graph and most perturbed genes (nodes). These nodes can be selected for training the model than earlier gene expression data based model which solely relied on the deviation in expression concerning healthy tissues. In Electrical circuits which is a rule-based graph of voltage conservation and current flow (Kirchhoff's voltage and current law) solve the similar problem where each junction can have distinct demand of current but overall supply of current needs synchronization with voltage. The Kirchhoff's law fits genomics data in two ways, firstly like electrical circuits where at each junction there is a loss and gain of current **I**. In genomics data that loss of connections in terms on molecular interacts is also multi-dimensional. For example, at each node of knowledge graph which is a gene, there could be loss or gain of mutations, these mutations can be shorter or multiple mutations such as CNV. Now, like electrical circuits even in genomics data, traversal to the next step will also depend on the status of the current node. We have modeled this using knowledge graph in genomics data. Second issue is the genomics growing field and the data distributed in silos across multiple datasets in multiple types, and luxuriant growing and not able to cover the complete human genome and complete data for each molecular subtypes till date. Such distributed growing of data makes the need to have data completion, now, Kirchhoff's second law which is a voltage law and based on traversal, it predicts the over the overall voltage. Similarly, in genomics knowledge graph, if data for few genes after integration of database is incomplete, applying the same voltage law helps to obtain completeness. The reason for completeness here is that at each node while traversing the node checks, addition or deletion of any knowledge in the previous node and ranks them based on the importance of data type decided by the user. Let us assume while traversing the knowledge graph we found a very well known cancer gene *BRCA1* where we have most of the data available for example, it has known

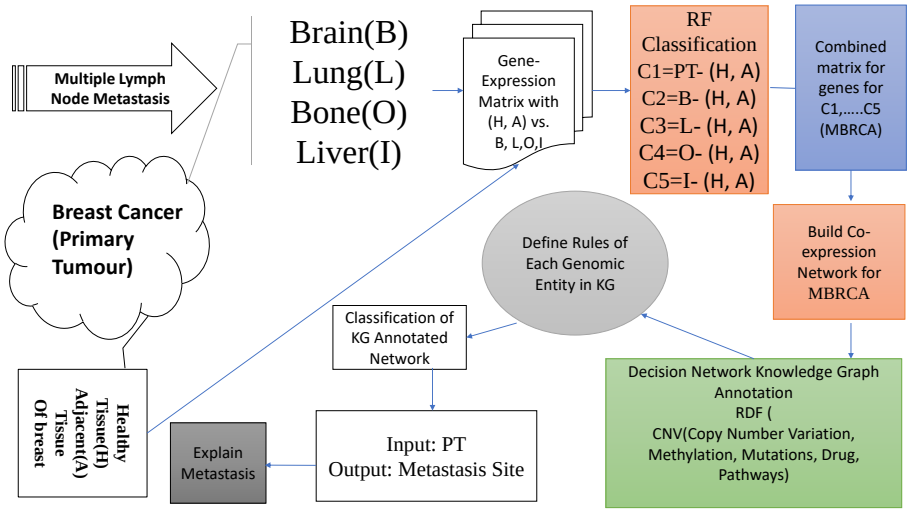


Figure 1: The Flow diagram of overall study for MBRCA bio-marker identification

disease, miRNA, CNV(start-end), drug information. It has a nearest neighbor a pseudo-gene which is called *ST6GALNAC4*. *ST6GALNAC4* is a gene which has minimum data available, now based on the traversal information from *BRCA1* stored in the index we can see if *ST6GALNAC4* falls in the same genetic range, if so, can they have same methylation status? Such domain specific rules helped to predict about *ST6GALNAC4* and allowed us to obtain data completeness based on the context of neighboring subject predicates. The overall workflow of our approach is shown in Figure 1.

Lastly in genomics data, biomarker discovery has a goal to predict the patient status, if the patient is going to relapse or not. Currently, this has been achieved through gene expression data which is very noisy and not deterministic. Alone in breast cancer, there are more than 500 biomarkers in almost two years with AUC above 80%. The issue with the gene expression method is that the genomics cannot be driven by a single entity such as gene expression whereas a gene is driven by more than one genomic data-types unlikely to predict erroneous disease status. Our Knowledge graph covers almost 90 %

of molecular data types such as genomics, drugs, pathways, miRNA, clinical and normal tissues data. We used data in layers as hidden layers in graph convolutional neural network to predict the disease release. We did validate our outcome on five independent breast cancer datasets apart from training and reported the performance. We also validated the knowledge completeness and traversal from Kirchhoff's law with other rules-based models and traversal algorithms and reported the outcome in the results section. Kirchhoff's law driven tractable topology will make it easy to build a data-driven decision network for relapse predictions and other medical decisions [24].

## 3. Related work

*Genomics and Breast Cancer Prediction.* The systemic view usually consists of alteration in pathways either due to change in expression or epigenetic change. However, due to the heterogeneity of breast cancer, it is challenging to point out one reason for metastasis. Tumor heterogeneity in breast cancer can involve ER/PR/HER2 status or stage dependent micro-environment for metastasis [3]. Several studies in the past tried to understand the systems behind the metastatic behavior of tumors in breast cancer. Studies have shown that the mutation process is the same in primary and relapse in breast cancer however it disseminates late keeping the driver genes [4]. Meta-analysis of breast cancer has revealed specific genes such as COX2 is down-regulated in MBC compared to normal tissues and not down-regulated in primary tumors compared with normal breast tissue and may be used as a differential marker in MBC [5]. This indicates that there is an existing path between various metastatic sites, primary tumor, and normal tissues. The suggestions of the pathway involved such as p53, ER1, ERB-B2, TNF, and WNT suggested that these markers selected from the comparative study of MBC against normal and primary tumors. Studies have also suggested that signatures derived from whole tissues which include (all tumor and stromal cells) or profiling of metastasizing tumor cells, including circulating tumor cells (CTCs) and migratory–disseminating tumor cells within

the primary tumor explains the cell migration in metastasis [6]. Studies covering the multiple metastatic sites also help to understand the drug response and planning of chemotherapy. For example, everolimus, the mutational status of genes allows the identification of some genes was potentially correlating tumor response/resistance [8]. Another advantage of site-specific studies for MBCA patients is to establish clinically relevant biomarkers. The inconsistency among ER, PR and HER2 status is due to the heterogeneous behavior of MBRCA [9]. Identification of extracellular matrix (stromal) genes can help metastasis may genetically seeding primary tumors, distinct differences that could be exploited to improve disease control.

### *3.1. Knowledge Graph(KG) Creation: Building Cancer Decision Network*

Creation of a knowledge graph to use for learning and graph mining, we adapted the knowledge graph to be fit with Kirchhoff's graph traversal and graph convolutional neural network. In the process of achieving so, we have formally explained underlying concepts as given below.

**Definition 1.** ***Knowledge Graph*** *A RDF Knowledge graph $G$ is a pair of disjoints sets $(S, P, O)$ where $S$,$P$ and $O$ represents subject predicts and object respectively such that $O$ is a subset of $[S]$ and the elements of $O$ are unordered pair of elements of $V$, whereas $P$ represents the unordered list of finite relations between $S$ and $O$ which makes the graph $G$ a finite graph.*

The Knowledge Graph here as shown in 3 have been cataloged in Five layers namely *Hidden Layer1: Mutations*, *Hidden Layer2: Pathways*,*Hidden Layer 3: Drugs*,*Hidden Layer 4: Diseases* & *Hidden Layer 5: Diseases* as reactions. The integrated layered knowledge graph provided input from training in Kirchhoff's Graph Convolution Neural Network.

**Definition 2.** ***Decision Networks*** *the RDF genomic decision networks are undirected RDF graph $G$ where the relation between $S$ and $O$ represents relapse status of the patient based on data sets $D = D1, D2, ....., DN$, where embedded of $D$ to build $G$ follows the following model.*

*Let us assume based on [25] that knowledge graph $G$ has an outcome space $\mathbb{R}$ with the set of molecular data types as $\mathbb{R}$ and set of actions based on a behavior of node as $\mathcal{C}$ and type of cancer selections as $\mathcal{S}$. Based on this to allow knowledge graph $G$ to pick entities on each node as the random variable $z_y \in [0,1]$ and $z_a \in [0,1]$ to indicate that when the outcome of $y$ and action $s$ is not in $\mathbb{R}$.*

$$\begin{aligned}\lambda^*(t,y,a,z_y,z_a) = \quad & \underbrace{\lambda^*(p)^*(z_y,z_a/t)}_{[A]GenomicsKnowledgeGraph} \\ & \underbrace{p^*(y/t,z_y,s,o}_{[B]DiseaseModel} \\ & \underbrace{p^*(a/y,t,z_a,p}_{[D]Outcomemodel} \\ & \underbrace{p^*(a/y,t,z_a,o}_{[E]Datainclusionmodel} \quad \underbrace{p^*(a/y,t,z_a}_{[C]Actionmodel})\end{aligned} \tag{1}$$

In cancer, it is a multi-model system where with every node which is gene here disease changes its behavior. Change of this behavior is difficult to track down by static traversal or binary decision systems. Here we developed a Rule-based decision system where the enrichment score of each node decided its strength. Here we have extended the algorithm from [25] we have adapted it for data inclusion and enrichment along with Breast cancer as a disease model. The model will help during traversal to decide the credibility of a node.

As per Algorithm 1 Firstly we built the knowledge graph. The novel approach we introduced is that after each node traversal we kept an index for loss and gain of information. Since in cancer, each gene does go through the mutations. These mutations can be insertions and deletions after each in Insertion, and deletion node chooses its pathways of progression. We have modeled the path to fit this dynamic modeling.

**Definition 3.** ***Complete Knowledge Graph & Knowledge boosting*** *The knowledge graph $G$ will be considered complete only if it satisfies following assertions.*

**Algorithm 1** Knowledge Graph for Cancer Decision Networks

**function** KGCreationforDecision($G$,$D_i$)

**Input: A RDF co-expression Network** $G$ **& datasets** $D = D_1, D_2.....D_n$

**Output: Integrated Genomics KG** $G^{'}$**, Enrichment Index** $E$

**Return:** $\mu_0$ **if** $D_i \leq 0$

IF $D_i \geq 0$ and Subject($S(G_i) \geq S(D_i)$)

Then $G_i \leftarrow RootGraph$

Generate($G_i, I(TraversalInformationStorageIndex)$)

$S_{T^{,}} \rightarrow Match(S, S_{D_i})$

$S_{T^{''}} \rightarrow Match(S, O_{D_i})$

Calculate(I) Loss or gain of Information at node $O$

Update $I$

IF $S_{T^{,}} \rightarrow True$

I $\rightarrow$ Information Gain

ELSEIF I $\rightarrow$ Information Gain

$G \rightarrow S_{T^{,}} + S_{T^{''}} \pm I$

**end function**

**function** TripleAcceptance using Kirchhoff's Law($G$,$D_i$, I)

Sort(I)

If Corr($\delta(G) \geq D_i$)

And Motif Size $\vec{\Delta} \geq 3$

$Mark(G[i], true)$

$ElseMark(G[S, O])$

**end function**

***Assumption 1.*** *The knowledge graph $G$ has minimum $n$ distinct resources $U = U1..U2..U_n$ such that integration of $U$ such that $U1 \times U2 \times ..U_n$ yields $S$ and $D$ such that $S = U1, U2...U_n \in U1 \times U2 \times ..U_n$ : owlsameAs(y,t)*

***Assumption 2.*** *The integration of entities from different sources is only possible if each entity as a minimum similarity score of 2. Here similarity score is the availability of entities in a minimum of two sources.*

***Assumption 3.*** *The integration of entities from different sources is only possible if each entity has a minimum similarity score of 2. Here similarity score is the availability of entities in a minimum of two sources.*

***Assumption 4.*** *The rule of knowledge graph for knowledge graph building has been declared based on domain knowledge during data completion first layer of the filtering index after applying these rules. For example, in CNV data from TCGA-BRCA and Drug data in Drugbank. Here data follows unique name assumptions from $U1...U_n$*

**Definition 4.** ***Centrality of Knowledge Graph*** *the centrality of knowledge graph $G$ based on degree $d_n$ where $N$ in the number of graphs from $U_N$ where degree $d = \sum_{n=1}^{n=N} a^N$ where $a$ is number of connections for each node of a graph $G$*

**Definition 5.** ***Co-expression Network*** *The co-expression network of any two entities in genomics graph $G$. Assuming for subject $s$ which can be gene, drug or pathway with the expression profile $s_j$ across $N$ data sets. So the co-expression correlation for each $p$ can be defined as $s_{jN} = \mod (cor(max(p_i), max(p_N)$*

During the knowledge graph building as we took care of information loss and gain, the primary task here is knowledge enrichment or knowledge boosting. To achieve this task, we have created assumptions(Rules) for boosting where each *Subject* can be mapped to its associated other graph subjects based on the property of the neighbor node.

Similarly, the baseline of knowledge graph is co-expression network which is necessarily a pair was *subject(S)* associations and each pair have literals as a correlation value, t -stat and o - the value of pair. This correlation based strength further boosted with KG Enrichment.

### *3.2. Kirchhoff's Law Traversal for KG*

Kirchhoff's law [26, 27] applied to electrical circuit graphs. Kirchhoff's law is divided into two laws namely Kirchhoff's voltage law(KVL) and Kirchhoff's current law(KCL) applied regarding spanning tree for undirected graphs and edges marked as resistors. KVL says that total sum of voltage for a network circuit graph *"may be called as voltage knowledge graph"* is zero. Similarly, KCL says that the sum of currents around any motif("graphlet") of circuit graph is zero. Now, this graph worked on conservation at each node and based on conservation at each node it works as a decision network to decide the path of traversal. We are the first group since 1847 who are adopting this law for graph pattern mining and graph traversal regarding knowledge apart from electrical spanning trees. The formal definition of the adapted KCL and KVL for Genomics RDF knowledge graph and its adaption for knowledge graphs are explained below. The relation between KVL and KCL with the genomics knowledge graph is shown in Figure 2 . The strength and novelty of this paper is Kirchhoff's law use for traversal where the information index will not only set the priority for the node, it will also help in graph mining as shown in Figure 2 where for each node *BRCA1* in the traditional approach. It will not be building relation between *b,c,e,d* whereas it will build relation keep subject as constant between the entities.

We have formally explained the Kirchhoff's law and knowledge completeness in Definition 6,7 and 8 as shown in Algorithm 2 where at each traversal point the information index decides which path to take based on the strength of the node.

**Definition 6.** ***Kirchhoff's Voltage law for Knowledge Graphs*** *The given RDF Graph G with disjoint sets (S,P,O) and datasets* $D_i$*, where* $i = 1, ..., n$*,*

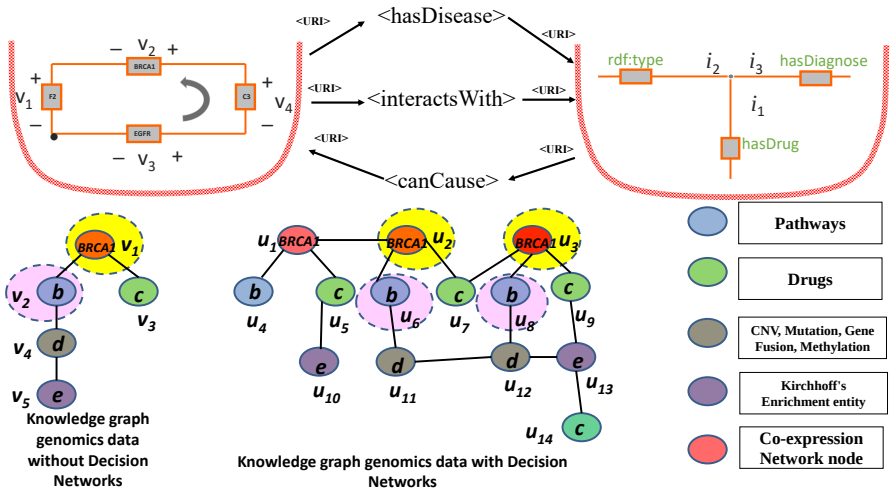


Figure 2: Implementation of Kirchhoff's law for Genomics knowledge graph

*node as $S$ and edge as $P$, Traversal as $T$, where total number of traversal can be defined as*

$$\sum_{n=1}^{P_j} T(D_i)T(G_i) \tag{2}$$

*. Considering that integrated knowledge graph $G$ has no repeated entities denoted as $\phi$. The KVL for Genomics knowledge graph can be defined as*

$$\sum_{n=1}^{P_j} T(D_i * \phi)T(G_i) = 0 \tag{3}$$

**Definition 7.** ***Kirchhoff's Current Law for Knowledge Graphs*** *The given RDF Graph $G$ with disjoint sets (S,P,O) and datasets $D_i$, where $i = 1, ..., n$, node as $S$ and edge as $P$ and a subgpah after each node traversal where if satisfies the set of rule $\mathcal{R} \implies G : (P, O)$ called as induced subgraph $G^{'}$. Then total current( Total data loss after integration) can be defined as*

$$< \Delta_i, G_j >= \sum_{n=1}^{P_j} \zeta(D_i * \phi)T(G_i^{'}) \tag{4}$$

*where $\zeta$ represents the overlapping knowledge or similar knowledge in knowledge graph $G$ also called as Isomorphism.*

**Definition 8.** ***Kirchhoff's Law for Knowledge Graph Traversal*** *The kirchoff traversal for a RDF knowledge graph $G$ which follows the rules defined as KVL and KCL for Genomics knowledge graph where each node is $S$ and edge is $P$ then if $S$ is subject of co-expression Graph $G$ and $S_i$ is subject of $G^{'}$ where $D_i \subseteq D_i$, where $i = 1, 2, ...S$, then at every node $S$ , $G \rightarrow G^{'} \cup D_i$ if and only if $P(G) = P(D_i)$ otherwise $G \rightarrow G^{'} \cap D_i$*

**Definition 9.** ***Kirchhoff's Law for Knowledge Graph Completeness*** *The knowledge graph $G$ is called complete knowledge graph is a pair of data sources in knowledge graph $D^{'}$ and $D^{''}$ represented as $G = (D^{'}, D^{''})$ and after applying to Kirchhoff's Traversal if $G \subset D^{'}$ and $G \subset D^{''}$ which results into the following equation.*

$$\sum G(Predicates\mathcal{P}) \leq \sum D^{'}(Predicates\mathcal{P}') + D^{''}(Predicates\mathcal{P}'') \quad (5)$$

*So Kirchhoff's completeness will be defined for Genomics knowledge graphs as given below. Let us assume the completeness on each node while applying the KVL is defined using cardinality by fixing subject and predicates and count the number of facts enriched in knowledge graph $\mathcal{G}$*

***Example 3.1.*** *As shown in Figure 2 Let's assume for this KG rules Kirchhoff's Rule $\mathcal{K}_\infty$ to $\mathcal{K}_\backslash$ can be realized as $\mathcal{K}_\infty : hasDisease(a, b) \leftarrow interectsWith(a, e) + hasDisease(a, d)$ and $\mathcal{K}_\backslash : canCauseDisease(a, b) \leftarrow interectsWith(a, u_1) + hasDisease(a, u2)$. Here the completeness assumption calculated by adapting Kirchhoff's law using [28].*

### *3.3. Classification using GCNN and KG (Cancer Decision Network)*

**Definition 10.** ***Motifs in Knowledge Graphs*** *For an undirected RDF Graph $G(S, P, O)$ where $S \subset O$ and can be denoted as $S[O^{'}]$ the the isomorphism of $G$ denoted as $G^{'}$*

***Example 3.2.*** *For a given knowledge graph $G_1$,$G_2$,$G_3$ triple patterns presented as $(S1, P1, O1), (S2, P2, O2), (S3, P3, O3)$ then $G1 \times G2 \times G3 < -(S1, O1), (O1, S2), (S2, O2)$ denoted as a motif or graphlet.*

**Algorithm 2** Kirchhoff's Law for KG Traversal and completeness

**procedure** KVLKG(Loss of Information($I_L$,Loss of Information($I_G$) for Particular node)
    **Input: An Integrated KG** $G$
    **Output: Integrated Genomics KG** $G^{'}$**, Enrichment Index** $E$
    $S \rightarrow G(S_i)$
    **for** each integer $i$ in $G$ **do**
        $\sum G(S_i) \rightarrow f(G(s,p,o)) = 0$
    **end for**
    Return $G^i$
**end procedure**
**procedure** KCLKG($I_G$,$I_L$ at Subnetwork)
    $S \rightarrow G(P_i)$
    **for** each integer $i$ in $G$ **do**
        $\sum G(P_i) \rightarrow f(G(p)) = 0$
    **end for**
    Return $G_{P'}$
**end procedure**
**procedure** KGTraversalusingKirchhoff(Rule Filtered KG $G_R$)
    **Input:** $G^i, G_{P'}$
    **Output: Rule Filtered KG** $G_R$ **and Infomation Loss & Gain index** $E$
    $sum = 0$
    **for** each subject $S$ in $G$ **do**
        $G(S,O) \rightarrow G(S) + G(O)$
        Loss Index $E_L \rightarrow G(S,O) - G(S)$
        **for** each Loss Index $E_L inG$ **do** $\sum E_L * G(P) = 0$
        **end for**
        Gain Index $E_I \rightarrow G(S,O) + G(S)$
        **for** each Loss Index $E_G inG$ **do** $\sum E_L / G(P) = 0$
        **end for**
    **end for**
    Return $sum$
**end procedure**

The compelling second argument in this paper is variable selection for training using a knowledge graph. We have used the graph topology approach. Based on the node strength the node which builds motifs (the group of nodes with higher strength) has been selected as the best features to train the convolution neural network. For the simplicity and as per the domain recommendations where triangular motifs represent the stability [29] We identified the triangles with the highest strength and used it for training. The motif counting followed Algorithm 3 . This is best of our knowledge, the only approach which uses graph topology and knowledge enrichment and knowledge strength for training.

**Definition 11.** ***Pattern Similarity in Knowledge Graphs*** *:For given knowledge Graphs $G^{'}$ and $G^{''}$ where $G, \cup(G_{'}, G_{''})$ the co-occurrence of motif $\mathcal{M}$ such that for Example 32*

*$S1 = O1, S1 = S2, S2 = O2, S2 = S1$ then motifs called as similar motifs or motif co-occurrence.*

### *3.4. Classification Using Kirchhoff's Graph Convolution Neural Network*

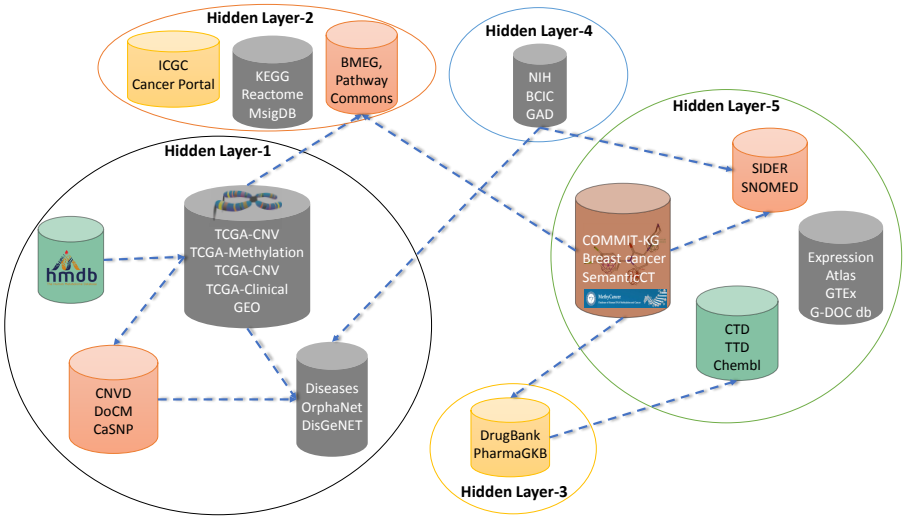


Figure 3: Genomics Knowledge Graph cataloged for 5 hidden layers for training

**Algorithm 3** Motif Identification from Knowledge Graph

```
procedure MotifCount(M)
    Input:RDF co-expression Network with Predicates P and objects O
    Output: Triangular motifs M with closed loop relation of (s, p, o)
    Subject S ← S'_i (Subject of KG-2), counts ← Counter(default=0)
    for O = 1.....G_P do:
        while doS_1 + Kirchhoff(L) < S_{1+O}
            Counts − −, S + +
            return Counts, Kirchhoff(index)
        end while
    end for
end procedure
procedure D(e)crementCounts(G,Kirchhoffindex(I))
    Counts[G(I) − −]
    for ( dosuffix in counts[index] ≤ G_P − 1 ))
        Counts[concate(I, suffix)] ≥ counts[suffix]
    end for
end procedure
procedure I(n)crementCounts(G,Kirchhoffindex(I))
    for ( doprefix in counts[index] ≤ G_P − 1 ))
        Counts[concate(prefix, I)] ≥ counts[Prefix]
        counts[G] ++
    end for
end procedure
procedure F(i)lterSizeThreeMotif(G,Kirchhoffindex(I))
    for C doounts in G
        counts[G] ++
        Count(I_pre,I_Post,I_Current)
        Push(Counts,I)
    end for
end procedure
```

**Definition 12.** ***Graph Convolution Neural Network*** *The given Knowledge Graph $G$ and hidden layer also called as "filter parameters" denoted as*

$$\sum_{n=1}^{5} F^n$$

*, then , overlapping filter parameters such that $G \subset F^n$, then learn the features $\mathcal{F}$ such that it satisfied following condition.*

***Assumption 5.*** *Let's assume the feature descriptor is motif denoted as $\vec{\Delta}$, where for every node $P$ and complete knowledge graph $G^c$ $\vec{\Delta}$ is such that, $\vec{\Delta}(S, P, O) \subseteqq G^c$ and $corr\vec{\Delta} = max(G^c)$. Then List of features from $G^c$. Then network layer based on $G^c$ can be written as $G^(l + 1) = f(G, \vec{\Delta})$.*

**Definition 13.** *For the given motif $\vec{\Delta}$, where $G \subset G^c$ with $cor(max(\vec{\Delta}))$ was annotated as kernel for graph convolution neural network.*

Once we have determined the motifs(best features) based on knowledge graph, we have trained the model using Graph convolution neural network on those features. We used only those motifs whose degree centrality is higher than means and Knowledge enrichment index is above 4 which means knowledge integration from at least four datasets. Further, we applied a graph convolution neural network on these filtered motifs based knowledge graph. We have extended Weissmuller-Lehman algorithm wherein Graph convolution neural network it has the kernel in images which keep moving around to find the best pixel point. We adapted this for topology motif as shown in algorithm 4 this moving kernel tries to find best motifs based on rewired knowledge enrichment within 702 nodes(genes). We apply the cut off here where we considered only complete nodes for selection. The criteria for completeness is already defined in Definition 9 and then trained the model using five hidden layers as shown in Figure 3. This approach has provided 11 such motifs which means 33 new genes which we call as the biomarker. We have validated and tested the performance of these variables as explained in the Results section.

**Algorithm 4** GCNN for Kirchhoff's Knowledge Graph extending Weissmuller-Lehman algorithm(1964)

**procedure** MOTIFGCNN($G$)

**Input:RDF KG** $G$**, motif Size**$M$**= 3, Kirchhoff's Rule index** $\mathcal{R}$

**Output: Features List** $\mathcal{F}$ **and Enrichment index** $E$

Subject $S \rightarrow P$ & $S \leftarrow 0$

**Repeat**

**for** $s_i \in$S **do**

$$G_i^{(i+1)} \leftarrow \sigma(\sum_{j=\mathcal{N}_)} 1/P_{ij} * G_j^P * W_1^P)$$

**end for**

Return $S++$, $P--$

**end procedure**

### *3.5. Prediction of Metastatic Site from Primary Tumour applying GCNN, KG, Cancer Decision Networks*

*Knowledge graph and convolution neural network.* As shown in Figure 4, we demonstrated the learning and prediction approach to predict the site of metastasis in breast cancer after obtaining an integrated knowledge graph. We have extended this work from our earlier approach where we used graph convolution neural network (GCNN) to predict relapse status in breast cancer [30]. However, our earlier work focused on the binary classification of breast cancer (Relapse/No-relapse). In this approach, the classification method is different since we have multiple sites of metastasis from the primary tumor. Further, this approach is beyond just prediction since it has to predict how much metastasis is happening on a given site from the primary tumor. Thus, the percentage of prediction depends on the molecular signature of each patient which is based on perturbations in gene expression and functional annotation of knowledge graph we developed. This way we will also be contributing by identifying the distribution of metastatic site based on range and deviation in expression. As

shown in Figure 4, we converted gene expression matrix into co-expression network where p-value for correlation (p-value¡ 0.01). The co-expression network paired the relation between two entities with the corresponding p-value of relation, correlation value, and t-static values. Now, expression network is enriched by Knowledge graph developed based on Algorithm 1. Further, we used Algorithm 2 for enrichment and integration of knowledge graph and co-expression network. Algorithm 2 explains that for each node in the co-expression network as mentioned earlier algorithm checks for all possible nearest neighbor. It attached all annotated from a direct and indirect connection to the current node. Further, it weighs the completeness at that moment and attaches with the node. Algorithm 2 clearly explains the rank weight of each node changes all nodes not visited for the given node. By Kirchhoff's traversal for every new connection established adds into the enrichment of the current node as well as all nodes having a relationship with the current node. As shown in Figure 4 the datasets are divided into 5 hidden layers. All these datasets have been manually grouped in the various hidden layers based on its relevance. Figure 3 explains the provenance for datasets where data sets having patient expression data are in one group, all drug databases are in on, all existing integrated graphs for genomics are in Layer 2. Thus, the current method also allows users to set the priority of genomic events for ranking based on the hypothesis. In current hypothesis, since we are looking into metastatic tumors while weighing the nodes during traversal, we will be giving a score (1) to gene expression, 0.5 to methylation, 0.25 to mutations, 0.125 to Drugs. This can be altered for other diseases. Once we have boosted co-expression network and knowledge graph as shown in Figure 4 we applied Algorithm 3 where we looked into the topological aspect of integrated knowledge graph. The reason behind this is that instead of using just gene expression data (numerical data) for feature selection we used the method where we looked into the motifs (size=3 to 7). It is known that in disease progression the bigger "hubs" from healthy tissues break into smaller motifs of various sizes. This represents the beginning of metastasis in breast cancer. As demonstrated in Figure 3 after obtaining enriched co-expression network we normalized the

network based on the ranks after traversal. After that, we applied the Algorithm 3 to identify motifs. Here motifs are the smaller graphlets of sub-networks, and higher frequency of these motifs defines the stability of a network. Usually, these motifs are higher in cancer stages which means cancer networks are less connected and unstable. As shown in 3 during traversal using Kirchhoff's law at each node we look for connected motifs and flag non-connected ones. Secondly the motif in Figure 3 shows three different color which means the node has been rewired from just being motifs of the only blue (gene expression motifs) to Drug(green) and Pathways. Motifs generated through enrichment from two different layers and tracking rewiring of knowledge graph nodes. The gene with no annotation flagged as "r'." In the second layer of convolution "r'" has been enriched based on neighborhood genes enrichment and two triangular motifs emerged. This is shown in Figure 3 as convolution layer 2. Layer after complete traversal based on five layers of knowledge graph from Figure 3 we stored all "r'." The assumption is such that all "r'" are the genes responsible for the perturbations in the larger hub of healthy breast cancer patients into smaller motifs and lead to cancer. The list of these genes which will be used as top features retrieved through graph pattern mining and ranked based on Kirchhoff's traversal is shown in Table 4.Selection of top variables based on network perturbations Kirchhoff's traversal is used first time on knowledge graph of any type of graph best of our knowledge till data. Now taking these genes (features), we trained the system and reported AUC(Area Under Curve) and other learning aspects as discussed in the Results section.

Once the model was trained using the features from Kirchhoff's approach the individual site of metastasis modeled for prediction. Now, here assumption is that through top features shows the pattern of expression in overall disease individual sites of metastasis is affected by a combination of the subset of the genes from top features. One more point to ponder here is that all the metastasis sites are some percentage of the primary tumor. Thus, as shown in figure 3 P1–P5 shows the percentage metastasis from the primary tumor. This shows the uniqueness of this model where the goal is not to predict the occurrences

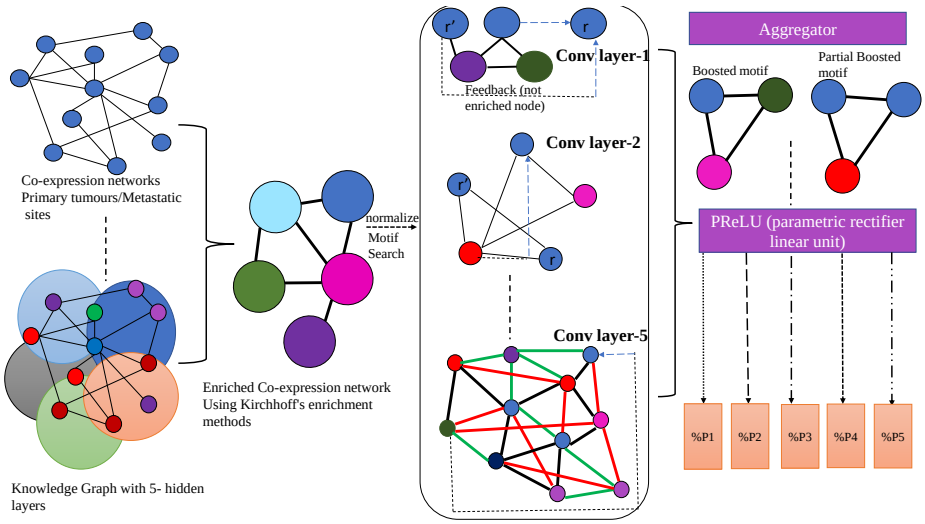


Figure 4: Multi-KG, Multi-class learning and distribution

but percentage contribution of each occurrence. The reason behind this is, in cancer progression sometimes lower percentage of metastasis in specific organs such as the brain, has a higher impact than a higher percentage of metastasis in the liver or bone. To predict this percentage and site-specific top features we gathered gene expression data from primary tumors and healthy and adjacent healthy tissues along with four metastatic sites of breast cancer namely Brain(B), Lung(L), Bone(O) and Liver (I). Further, we have paired primary tumor along with B, L, O and I with Healthy(H) and Adjacent Healthy(A) and prepared gene expression matrix. We did random forest classification of these matrices and yielded *Gini* top variables. Further, we have selected variables for each classification. To impede possible errors from current differential expression analysis method, we interposed classification importance scores as differential molecular property. After the selection of best variables, we have merged the result and based on overlapping genes we have build co-expression network using classification scores. After that, we did the clustering of the genes. After clustering, we did disease enrichment analysis(DEA) of each cluster. Then we

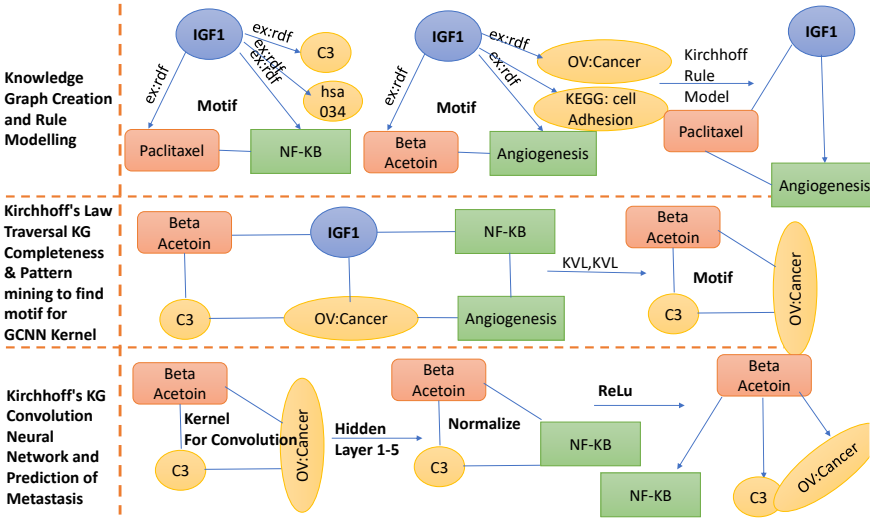


Figure 5: The Flow diagram of overall study for MBRCA bio-marker identification

looked into the disease from each cluster to see represented disease and phenotypes retrieved from cluster built from primary tumors and metastasis sites. Further, we searched for any existing path among the clusters. Genes are available in all these paths called breast cancer metastatic sites. We looked into associated pathways, GO enrichment and survival analysis of these genes and established metastatic biomarkers for breast cancer, not just based on a single site of expression but from all sites of expression.

## 4. Results & Discussion: Biomarkers, Diagnostic, Prognostic evaluation and Metastatic site Prediction

We designed the experiment in multiple steps to evaluate and predict result such as AUC (Training and Testing) of top features, comparison with other graphs which is similar to knowledge graph for genomics and relevance of genes retrieved in breast cancer progression.

### *4.1. Performance of Model Training and & Testing*

. We have trained the model using datasets retrieved through TCGA and GEO. The detailed list of samples in each group for each data type is shown in Table 2. We have normalized the data and removed the batch effects using the Combat tool. Combat models the batch effects by including them in the design matrix and modeling the unadjusted data. Removal of the batch effect

Table 1: GCNN-Kirchhoff's Training and Validation

| Algorithm | Training TCGA-BRCA | | Validation-GSE47561 | |
|---|---|---|---|---|
| - | AUC | P-value | AUC | P-value |
| NN | 71[CI 68-77.8] | 0.023 | 58 [CI 55-61] | 0.35 |
| RF | 67 [CI 64.5-69.99] | 0.047 | 75[CI 68.8- 79.8] | 0.025 |
| SVM | 70[CI 69.12-71.14] | 0.054 | 63 [CI 58.8- 66.8] | 0.0057 |
| GCNN-Kirchhoff | 83.8 [CI 81.7-84.84] | 0.048 | 79.85 [CI 77.56- 81.94] | 0.037 |

ensures that the linear model accurately accounts for the degrees of freedom associated with modeling the batch effect. Further after getting the list of genes using Kirchhoff's method as shown in Table 4 further the list of genes and its hazard ration is also shown. The training Results are shown in Table 1. The training results show that AUC (area under the curve) with NN(Neural network), RF(Random Forest) and SVM (Support vector machine) are 71, 67 and 70 respectively. However, with our method GCNN-Kirchhoff's, we could obtain 83.8 %.

The point to ponder here is that in multi-class classification AUC, we did not combine all data. However, after training and acquiring features, we combined all data from Table 2 to predict the site as reported in Table 5. The performance was consistent with validation and test datasets. For instance, as shown in Table 1 with the validation dataset (GSE47561) the performance of GCNN-Kirchhoff is 79% edging ahead of RF, NN, and SVM. We also tested the performance of top features retrieved as shown in Table 4. The performance of test data is shown in Table 3. We observed in Table 3 that the performance of GCNN-Kirchhoff in test data also outperforms NN, RF, and SVM. The performance using GCNN-

| **Dataset** | **Normal** | **Brain** | **Lung** | **Bone** | **Liver** | **Primary Tumor** |
|---|---|---|---|---|---|---|
| TCGA-BRCA | 2 | 0 | 2 | 4 | 0 | 1090 |
| GSE12276 | 0 | 204 | 0 | 0 | 0 | 0 |
| GSE39494 | 0 | 0 | 0 | 5 | 0 | 5 |
| GSE46928 | 0 | 11 | 0 | 0 | 0 | 0 |
| GSE50567 | 6 | 0 | 0 | 0 | 0 | 0 |
| GSE4611 | 0 | 0 | 0 | 0 | 0 | 218 |
| GSE39494 | 0 | 0 | 0 | 5 | 0 | 5 |
| GSE2294 | 0 | 0 | 0 | 0 | 0 | 96 |
| GSE27830 | 0 | 0 | 0 | 0 | 0 | 155 |
| GSE14017 | 0 | 0 | 4 | 25 | 0 | 0 |
| GSE14020 | 0 | 22 | 20 | 18 | 5 | 0 |
| GSE10893 | 14 | 9 | 12 | 0 | 0 | 275 |
| GSE14682 | 0 | 60 | 0 | 0 | 0 | 0 |
| GSE32489 | 0 | 0 | 5 | 0 | 7 | 0 |
| GSE3521 | 0 | 5 | 4 | 0 | 1 | 4 |
| GSE38057 | 0 | 41 | 0 | 0 | 0 | 46 |
| GSE46141 | 0 | 0 | 2 | 5 | 16 | 11 |
| GSE46928 | 0 | 11 | 0 | 0 | 0 | 41 |
| GSE5327 | 0 | 0 | 11 | 0 | 0 | 47 |
| GSE54323 | 0 | 0 | 0 | 7 | 4 | 0 |
| GSE56493 | 0 | 0 | 0 | 5 | 27 | 0 |
| GSE58708 | 0 | 0 | 0 | 0 | 3 | 0 |
| GSE61723 | 8 | 0 | 0 | 0 | 0 | 13 |

Table 2: The compendium of datasets for primary and metastasis sites (Brain, Lung, Bone and Liver) along with Normal

Table 3: GCNN- Kirchhoff's KG performance table-Testing. (*Confidence Interval [CI])

| Algorithm | GSE20685 | | GSE25055 | | GSE22219 | | GSE12276 | |
|---|---|---|---|---|---|---|---|---|
| - | AUC | P-value | AUC | P-value | AUC | P-value | AUC | P-value |
| NN | 74.2 [73.8-78.7]* | 0.029 | 66.5 [64.8-67.8]* | 0.4 | 62.7 [54.9-64.3]* | 0.07 | 57 [51.8-59.6]* | 0.047 |
| RF | 76 [74.2-80.5]* | 0.011 | 67 [63.8-72.9]* | 0.70 | 66 [65.8-69.11]* | 0.026 | 74 [67.9-78.2]* | 0.32 |
| SVM | 63 [57.5.2- 64.9]* | 0.36 | 54 [51.3-59.6]* | 0.02 | 68 [64.8-72.3]* | 0.022 | 61 [60-68.4]* | 0.037 |
| GCNN-Kirchhoff | 81 [80.6-87.4]* | 0.051 | 78 [77.2-82.5]* | 0.029 | 84 [81.9-89.3]* | 0.006 | 76 [70.8-77.4]* | 0.004 |

Kirchhoff ranges from minimum 76 to the maximum 84 in test data. Further, we obtained the performance and survival of 32 genes from Table 4 and the performance of genes associated with each tissue type. The median survival time for each group is between 120 to 125 days from the primary tumor to each metastasis site. Significant p-value in each group and low hazard ration (¡1) suggests the association of genes with smaller risk group. Thus, they lead to metastasis. These genes need to be further investigated by drug discovery units for clinical relevance. Finally, we gave patient data from TCGA a primary tumor and predicted the metastasis site based on genes we retrieved through our motif and Kirchhoff's approach. The Results is for 2006 primary tumor patients are shown in Table 5. We tested the metastasis site for TCGA patients in four categories. The first category is where we looked into the single site only from primary tissue (PT). The AUC of prediction of metastasis site is 77, 68, 56 and 87 for Bone, Lung, Liver, and Brain respectively. The trend shows that in single site prediction, our approach is performing better with Bone and Brain. Further, it elucidated that metastasis is trenched from the primary tumor.

Here, from primary tumour 8%, 15%,17%, and 2.1%. We also calculated the F1 Score 87, 56, 59 and 91 for Bone, Lung, Liver, and Brain respectively. The p-value for a single site prediction is 0.0361. Thus, it shows that from a primary tumor to enriched metastasis site are Bone and Brain. Further, we looked into two site prediction from our model. Here, as shown in Table 5 Bone+Brain, Lung+Liver, and Lung+Brain have 88, 89 and 75 predictions

AUC and Lung+Liver shared 8.96% of Primary Tumour. F1 score is also significant in all groups to accept the Bone+Lung group. The p-value of two site prediction is 0.0000325. With three site prediction, Bone+Liver+Brain and Liver+Lung+Brain have 78 and 77 AUC. Liver+Lung+Brain Shares 22.86 of Primary Tumour. The p-value in case of three site prediction is 0.00012. Finally, we applied our model for all four sites Brain+Lung+Liver+Brain and achieved 84% accuracy. It shares 56.8% of Primary Tumor with F1-Score of 0.87% and p-value 0.044. This is one of its kind of predictor predicting metastasis from primary tumor not just using gene expression data but multi-omics data with an accuracy of $> 80\%$. Here, we did face the issue of sample imbalance as we can see in Table 4 and 5 where the primary tumor has 2006 samples and a minimum 60 samples for Liver. We used randomization to remove the sample bias while training and avoid class imbalance. In the next section, we discussed the evaluation of the model.

Table 4: Comparison of Genes within metastatic sites (H.R.=Hazard Ratio, MST=Median Survival Time)

| **Sites** | **Kirchoff's GCNN(PT)** | **Brain(B)** | **Bone(O)** | **Lung(L)** | **Liver (I)** |
|---|---|---|---|---|---|
| **Genes** | ABCC5<br>DPP3<br>TP53<br>AP2S1<br>ZNF703<br>RYR2<br>CDH11<br>TSPAN5<br>TSPAN1<br>NCOR1<br>FLT1<br>MYH7<br>SEC23B<br>IDH3B<br>ARL4C<br>RASGRP1<br>MKI67<br>OPTN<br>RAB11A<br>MLPH<br>ADCY3<br>ENPP1<br>TMEM<br>GNB2<br>GNG4<br>SH3GLB1<br>COPE<br>NFKB2<br>CCR7<br>PTEN<br>NHERF1<br>F3 | SH3GLB1<br>COPE<br>MLPH<br>NCOR1<br>DPP3<br>F3<br>CCR7<br>ARL4C<br>OPTN<br>ADCY3<br>SEC23B<br>NHERF1<br>TP53<br>GNB2<br>COPE<br>RAB11A | ABCC5<br>SEC23B<br>NHERF1<br>TSPAN1<br>RYR2<br>F3<br>NFKB2<br>ENPP1 | ABCC5<br>DPP3<br>TP53<br>RYR2<br>FLT1<br>MYH7<br>IDH3B<br>ARL4C<br>RASGRP1<br>MKI67<br>RAB11A<br>MLPH<br>GNG4<br>COPE<br>NFKB2<br>NHERF1<br>F3 | TP53<br>AP2S1<br>ZNF703<br>RYR2<br>TSPAN5<br>TSPAN1<br>NCOR1<br>FLT1<br>SEC23B<br>IDH3B<br>ARL4C<br>MKI67<br>OPTN<br>RAB11A<br>MLPH<br>ADCY3<br>TMEM<br>GNB2<br>GNG4<br>SH3GLB1<br>COPE<br>NFKB2<br>CCR7<br>PTEN<br>NHERF1<br>F3 |
| **Site H.R.** | **0.64[CI 0.55-0.75]** | 0.66[CI 0.57-0.77] | 0.75[CI 0.64-55] | 0.59[CI 0.50-0.69] | 0.65[CI 0.55-0.75] |
| **MST-Months** | 126 | 123 | 121.5 | 121 | 127 |
| **P-value** | 2.1e-8 | 2.0e-7 | 0.0003 | 1.7e-11 | 3.0e-8 |
| **Samples** | 2006 | 349 | 74 | 63 | 60 |
| **Total Gene** | 32 | 16 | 08 | 17 | 26 |

Table 5: Prediction using Kirchooff's GCNN in Breast Cancer Patients)

| Metastatic Site | % AUC Site Prediction | % Metastasis(% of PT) | p-value | F1 Score |
|---|---|---|---|---|
| Single Site | | | 0.0316 | |
| Bone | 0.77 | 8 | | 0.87 |
| Lung | 0.68 | 15 | | 0.56 |
| Liver | 0.56 | 17 | | 0.59 |
| Brain | 0.87 | 2.1 | | 0.91 |
| Two Sites | | | 0.0000325 | |
| Bone+Lung | 0.62 | 3.7 | | 0.27 |
| Bone+Liver | 0.56 | 2.8 | | 0.65 |
| Bone+Brain | 0.88 | 0.59 | | 0.48 |
| Lung+Liver | 0.89 | 8.91 | | 0.66 |
| Lung+Brain | 0.75 | 1.12 | | 0.62 |
| Liver+Brain | 0.69 | 2.86 | | 0.67 |
| Three Sites | | | 0.00012 | |
| Bone+Lung+Liver | 0.52 | 18.7 | | 0.17 |
| Bone+Lung+Brain | 0.41 | 12.4 | | 0.14 |
| Bone+Liver+Brain | 0.78 | 9.94 | | 0.83 |
| Liver+Lung+Brain | 0.77 | 22.86 | | 0.71 |
| Four sites | | | 0.044 | |
| Bone+Lung+Liver+Brain | 0.84 | 56.8 | | 0.87 |

## 5. Evaluation

### *5.1. Model Evaluation*

As discussed in the result section, we evaluated the performance of our model on four independent datasets as shown in Table 3. We evaluated the performance on GEO datasets GSE 20685, GSE 25055, GSE22219 and GSE12276. We compared our performance with Neural Network, Random Forest and Support Vector Machine. The AUC is 81, 78, 84 and 76 in GSE 20685, GSE 25055, GSE22219 and GSE12276 respectively. The model which is performing near to our model is NN in GSE260685 and GSE12276 and RF in GSE20685 and GSE12276. Thus, our model is outperforming all other models even on independent test datasets.

### *5.2. Knowledge Graph Evaluation*

We have also validated the KG(Knowledge Graph) in terms of regarding knowledge boosting and functionality. As shown in Table 6 other integrated genomics datasets/knowledge graph such as BMEG, GISGeNET and CBioPortal basically has 8, 14 and 1 datasets. We did the knowledge enrichment by incorporating 36 datasets as shown in Figure 3. Further, we did rule-based integration and cataloging of data based on molecular signatures. Thus, it is the first knowledge graph which is rule-driven, and a rule is modeled using an electrical engineering algorithm adapted for graph called Kirchhoff's law makes it better than other algorithms and novel.

Table 6: Comparison of Kirchhoff's KG with other integrated genomic graphs

| **Knowledge Graphs** | **Datasets** | **Rule Based** | **Domain Driven** | **Knowledge Enrichemnt** |
|---|---|---|---|---|
| BMEG | 8 | No | Yes | No |
| DisGeNET | 14 | No | No | Partial |
| cBioPortal | 1 | No | Yes | No |
| Kirchhoff's | 36 | Yes | Yes | Yes |

### 5.3. Kirchhoff's Law Evaluation

We also evaluated the Kirchhoff's law with another traversal algorithm, though this comparison is not fair since there is no state of the art where Kirchhoff's law has been used for traversal and rule-based decision network. Here, Kirchhoff's law is to take more time as shown in Table 7 due to its index-based approach of knowledge conservation. Here in this traversal, each node gets enriched every time and discards noise from the data and resets the ranking based on domain and research requirements. Further, this is nor biased towards any datasets, so in future any inclusion and exclusion of datasets will be smooth which supports genomics and precision medicine where data is exponentially growing every year.

Table 7: Comparison of Kirchhoff's Traversal Algorithm with other traversal Algorithm. BFS=Breadth First Search, DFS=Depth First Search, RD=Reverse-Delete, FF=Ford-Fulkerson, CGI=Cancer Genomics Implementation

| | **Dijkstra's** | **BFS** | **DFS** | **Kruskal's** | **RD** | **FF** | **Kirchhoff's** |
|---|---|---|---|---|---|---|---|
| Context Aware | No | No | No | No | Yes | No | Yes |
| RDF Support | No | Yes | Yes | No | No | No | Yes |
| KG Boosting | No | No | No | No | No | No | Yes |
| Enrichment Index | No | Partial | Partial | No | No | No | Yes |
| Traversal Time (msec) | 0.21 | 0.28 | 0.17 | 1.21 | 0.015 | 0.56 | 2.57 |
| Predictive Traversal | No | No | No | No | No | No | Yes |
| CGI | No | Yes | Yes | No | No | No | Yes |

As shown in Table 7 we evaluated traversal part of Kirchhoff's approach with another traversal algorithm such as Dijkstra's, Breadth First Search, Depth First Search, Kruskal's, Reverse-Delete, Ford-Fulkerson. The parameters used for evaluation are Context Aware-means available implementation for multi-omics data, RDF support- means if the algorithm was tested on RDF data or not, KG Boosting- means if algorithm supports knowledge enrichment or not, Enrichment Index- here it checks if algorithm keeps track of enriched information or not. In the present research we build an enrichment index to to track the location of enrichment, since with every traversal our enrichment is changing Traversal time. To achieve this we obtained the nodes with highest and least

degree which are BRCA1 and PTEN then calculated the traversal time between these two nodes,Predictive traversal- where if algorithm can predict the knowledge based on nearest neighbor or not, in our approach after all traversal if node is missing some information we predict that enrichment based on likehood and similarity of nearest neighbor and lastly we looked that how many of these algorithms have implementation on cancer genomic data. The result clearly indicates that due to enrichment at each node and predictive traversal, our model is taking longer time. Otherwise it is outperforming all other traversal approaches.

### *5.4. Knowledge Enrichment Evaluation*

In Genomics, data enrichment is one of the critical factors due to the availability of data in silos across the web in multiple databases and rapidly growing data. We used the layered knowledge graph approach with Kirchhoff's traversal to integrate 36 types of genomics databased on its molecular subtypes as shown in Figure 3. We have compared the enrichment using our approach and current such datasets in terms of link boosting as shown in Table 8. The point encirclement links are shown in Table 8 is extracted keep the number of human genes 22000 till all neighbors for its genes.

We compared the various multi-omics studies as shown in Table 8. The First column shows the datasets or studies associated with multi-omics data. Since very few studies have followed the knowledge graph approach, we have listed them if they have a multi-omics integrated query based or graph based studies. As we can see with Kirchhoff's KG approach number of the link are higher than other similar approaches, keeping in mind that our approach is Rule-based.

Table 8: Knowledge enrichment comparison of Kirchhoff's KG with other integrated genomic graphs

| **Dataset** | **Type** | Links |
|---|---|---|
| BMEG | Integrated | 568745 |
| DisGenET | Integrated | 7758968 |
| Bio2RDF | Query Based | 258987586 |
| cBioPortal | Query Based-TCGA | 189567 |
| COMMIT-KGBC | Knowledge Graph | 192484 |
| LinkedOmics | Query Based | - |
| MODMatcher | Integrated-TCGA | 258896 |
| OASIS | Integrated | 568571 |
| Pathway Commons V2.0 | Pathways-Integrated | 42558 |
| ComPath | Pathways-Integrated | 77989 |
| Enrichr | Integrated | 89956 |
| DAVID | Pathways-Integrated | 109366 |
| Stanford TCGA-explorer | Integrated -TCGA | 44589 |
| Kirchhoff's KG | Knowledge Graph | 459844005 |

## 6. Future Work

Artificial intelligence(AI) and digital health are together going to play a vital role in medicine. However, the requirement is to bring clinical data, Patient data and genomics data for better treatment management in cancer. Especially in cancer, which is a combination of multiple diseases in terms of its stages, molecular subtypes, and data dimensionality. Any oncology system to improve prognosis in patient would require a knowledge graph in layers as in our case. Further, the layers knowledge graph will determine which type of molecular type or layer is driving cancer. Thus, this approach can be applied to any other cancer types. Secondly, using pattern mining to multi-model knowledge graph will add a new approach for feature selection. Thus, any disease which is multi-step such as cancer, diabetes, Alzheimer's disease, and viral disease such

as pattern mining will help to make a robust and mechanism driven approach. Further, Kirchhoff's electrical engineering approach into graph traversal, pattern mining and rule-based modeling will help big data medicine having a mechanism based data reduction, features selection and knowledge boosting. Further, such approaches will improve patient prognosis and survival in the future with better integration with FHIR and other clinical systems.

## 7. Conclusion

The present paper presents a novel approach of biomarker identification for predicting the site metastasis from primary tumors by the integration of multi-omics data called Genomics knowledge graph. It shows that the multi-omics data helps to predict the site of metastasis better by using Kirchhoff's law for traversal, pattern identification, feature selection, and multi-model prediction.

## 8. Funding

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.

## References

[1] Estimation of the Number of Women Living with Metastatic Breast Cancer in the United States Agela B. et al Prev June 1 2017 (26) (6) 809-815;

[2] Redig AJ, McAllister SS. Breast cancer as a systemic disease: a view of metastasis. Journal of internal medicine . doi:10.1111/joim.12084.

[3] Meng X et al. Receptor conversion in metastatic breast cancer: a prognosticator of survival. Oncotarget. 2016; doi:10.18632/oncotarget.12114.

[4] Yates, L. R., et al (2017). Genomic evolution of breast cancer metastasis and relapse. Cancer Cell, 32(2), 169-184.

[5] Bell, R. et al (2017). Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. Current molecular medicine, 17(3), 200-210.

[6] Karagiannis, et al (2016). Signatures of breast cancer metastasis at a glance. J Cell Sci, 129(9), 1751-1758.

[7] Barney, et al (2016). The predictive link between matrix and metastasis. Current opinion in chemical engineering, 11, 85-93.

[8] Omarini, C.,et al(2018). Mutational Profile of Metastatic Breast Cancer Tissue in Patients Treated with Exemestane Plus Everolimus. BioMed research international, 2018.

[9] Soria, J. C. at al (2012). Discrepancies between primary tumor and metastasis: a literature review on clinically established biomarkers. Critical reviews in oncology/hematology, 84(3), 301-313.

[10] Kimbung, S., et al (2015). Transcriptional profiling of breast cancer metastases identifies liver metastasis-selective genes associated with adverse outcome in luminal A primary breast cancer. Clinical Cancer Research, clincanres-0487.

[11] Hussein, O., & Komarova, S. V. (2011). Breast cancer at bone metastatic sites: recent discoveries and treatment targets. Journal of cell communication and signaling, 5(2), 85-99.

[12] Wu, J. M. at al (2008). Heterogeneity of breast cancer metastases: comparison of therapeutic target expression and promoter methylation between primary tumors and their multifocal metastases. Clinical cancer research, 14(7), 1938-1946.

[13] Naomoto, Y. at al(1999). Multiple Liver Metastases of Breast Cancer: Report of a Case Successfully Treated with Hormone—Cytokine—Chemotherapy. Japanese journal of clinical oncology, 29(8), 390-394.

[14] Leri, J. P. (2018). Metastatic Cancer of the Thoracic and Lumbar Spine Presenting as Mid-and Low Back Pain in a Long Distance Runner. Journal of Chiropractic Medicine, 17(2), 121-127.

[15] Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nature methods, 11(6), 599.

[16] Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118-127.

[17] Law, C. W. at al(2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology, 15(2), R29.

[18] Raftopoulou, P., & Petrakis, E. G. (2008, March). iCluster: a self-organizing overlay network for P2P information retrieval. In European Conference on Information Retrieval (pp. 65-76). Springer, Berlin, Heidelberg.

[19] Morris, at al (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC bioinformatics, 12(1), 436.

[20] Tang, Y. at (2015). CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems, 127, 67-72.

[21] Colaprico, at al(2015). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research, 44(8), e71-e71.

[22] Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 559.

[23] Lauria, M., Moyseos, P., & Priami, C. (2015). SCUDO: a tool for signature-based clustering of expression profiles. Nucleic acids research, 43(W1), W188-W192.

[24] Fan, C. Y., Chang, P. C., Lin, J. J., & Hsieh, J. C. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing, 11(1), 632-644.

[25] Schulam, P. and Saria, S., 2017. Reliable decision support using counterfactual models. In Advances in Neural Information Processing Systems (pp. 1697-1708).

[26] Kirchhoff, Gustav, and Robert Bunsen. "Chemische analyse durch spectralbeobachtungen." Annalen der Physik 186, no. 6 (1860): 161-189.

[27] Kirchhoff, G. (1847), Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. Ann. Phys., 148: 497-508. doi:10.1002/andp.18471481202

[28] Tanon, T.P.,at al, 2017, October. Completeness-Aware Rule Learning from Knowledge Graphs. In International Semantic Web Conference (pp. 507-525). Springer, Cham.

[29] Tran, N.H., Choi, K.P. and Zhang, L., 2013. Counting motifs in the human interactome. Nature communications, 4, p.2241.

[30] Jha, Alokkumar, et al. "Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.