Çizge tabanlı Biyomedikal Bilgi Keşfi
Graph-based Biomedical Knowledge Discovery

Osman Altuner[1], Burcu Bakir-Gungor[2], Gokhan Bakal[3]
Elektrik ve Bilgisayar Mühendisliği[1], Bilgisayar Mühendisliği[2,3]
Abdullah Gül Üniversitesi[1,2,3]
Kayseri, Türkiye
osman.altuner@agu.edu.tr[1], burcu.gungor@agu.edu.tr[2], gokhan.bakal@agu.edu.tr[3]

2024 32nd Signal Processing and Communications Applications Conference (SIU) | 979-8-3503-8896-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/SIU61531.2024.10600774

***Özetçe*—Dijitalleşme süreci tüm dünyada oldukça yüksek bir hızla ilerlemektedir. Bu durum günümüz yaşantısında bir çok kolaylık sağladığı gibi ortaya çıkan devasa dijital verilerin analizi ve işlenmesi gibi bir problemi de beraberinde getirmektedir. Bu durum yayınlanan akademik çalışmalar için de geçerlidir. Bu anlamda çalışmalar dahilinde bulunan yenilikçi bilgilere ulaşmak için her bir çalışmayı değerlendirme süreci oldukça zahmetli bir süreci gerektirmektedir. Bu sebeple yapılan bu çalışmada hedef hastalıklar özelinde elde edilmiş yayınlar metin analiz süreçleriyle analiz edilmiş ve anlamlı terimlerin biyomedikal ilişkiler üzerinden bağlanmasını sağlayan çizge yapısına dönüştürülmüştür. Elde edilen yoğun çizge yapısı üzerinde treats (tedavi edici), causes (sebep verici), associated_with (ilişkili) gibi önemli bağlantılara sahip ikili biyomedikal varlıklar sorgulanmıştır. Sorgu sonuçlarına göre elde edilen varlık ikilileri manuel arama yöntemiyle de teyit edilmiş ve gerçek bağlantılar olduğu ispatlanmıştır. Bu çalışmayla birlikte, bilinen biyomedikal varlıkların önerilen yaklaşımla elde edilmesi uzun zaman gerektiren manuel arama problemini çözmesi hedeflenmektedir. Ayrıca birden fazla ikili bağlantı örüntüleriyle bilinmeyen/keşfedilmemiş olası yeni ilişkiler (tedavi edici, sebep verici, ilişkili vb.) elde etme potansiyeli de bulunmaktadır.**

Anahtar Kelimeler — metin madenciliği; bilgi keşfi; çizge analizi; neo4j.

***Abstract*—The digitalization process is progressing at a very high speed all over the world. While this situation provides many conveniences in today's life, it also brings along a problem such as analyzing and processing the huge digital data. This also applies to published academic studies. In this sense, the process of evaluating each study to access previously unknown information within the studies requires a very laborious process. For this reason, in this study, the publications obtained for the target diseases were analyzed by text analysis processes and converted into a graph structure that enables the linking of meaningful terms through biomedical relationships. On the dense graph structure obtained, binary biomedical entities with important links such as treats, causes, associated_with were queried. The entity pairs obtained according to the query results were also confirmed by manual search method and proved to be real connections. In this study, retrieval of known biomedical entities with the proposed approach solved the time-consuming manual search problem. There is also the potential to obtain unknown/unexplored possible new relationships (e.g., therapeutic, causal, etc.) with multiple binary linking patterns.**

Keywords — text mining; knowledge discovery; graph analysis; neo4j.

979-8-3503-8896-1/24/$31.00 ©2024 IEEE

I. GIRIŞ

Güncel teknolojik gelişmeler hayatımızın her yönünü ciddi bir şekilde yeniden yapılandırmaktadır. Özellikle dijital dönüşüm kavramı, anlık ve zamandan tasarruf sağlayan günlük yaşam rutininde birçok avantajı beraberinde getiren önemli dönüşümlerden birisidir. Daha somut bir örnek vermek gerekirse, insanlar herhangi bir zaman kısıtlaması olmaksızın dünya çapında belirli bir konudaki son güncellemelere veya bilgilere kolayca anlık bir surette ulaşabilmektedir. Günlük yaşantımızda, dijital dönüşümün tam anlamıyla tamamlanmamış olmasına rağmen dijital ortamda bulunan metinsel verilerin boyutu günbegün hızla artmaktadır [1]. Bu verilere örnek olarak makaleler (akademik ve akademik olmayanlar dahil), klinik hasta kayıtları, web sayfaları, çevrimiçi ortamlarda yapılan alışveriş kayıtları ve sosyal medya (Twitter, Facebook ve Reddit gibi) paylaşımları verilebilir. Hatta giyilebilir akıllı saat gibi teknolojik cihazlar bile kullanıcısı hakkında bir çok veri tipinde kişisel bilgiler oluşturmaktadır [2].

Bu anlamda dünya genelinde akademik çalışmaların sayısı da her gün oldukça yüksek bir hızla artmaktadır. Bu denli bir artışa sahip veri kütlesinin manuel yordamlarla analiz edilmesi mümkün olmamaktadır. Bu sebeple hesaplamalı ve özellikle makine öğrenmesi temelli akıllı modellerden yoğun şekilde yararlanılmaktadır [3, 4]. Bu çalışmada da medikal akademik yayınların tutulduğu PubMed [5] açık erişim deposunda yer alan çalışmalar analiz edilmiştir. Yapılan analizler önceden belirlenen hedef hastalıklara ait medikal çalışmalar üzerinden gerçekleştirilmiştir. Her bir hastalık için elde edilen çalışmalar önişleme adımlarından geçirilip bağlamsal olarak veri setini yansıtan en önemli terimler yardımıyla spesifik ilişki türlerinin bulunduğu SemMedDB tripletleri kullanılarak hedef hastalık çizge yapıları Neo4j platformu [6] üzerinden elde edilmiştir. Sonrasında her bir çizge yapısından ilgili hastalık düğümünü içerecek ikili ilişki sorguları çalıştırılarak sonuçlar elde edilmiştir. Bulunan sonuçlar var olan ilişkileri teyit etmiş ve izlenen yolun doğruluğu desteklenmiştir.

Çalışmanın geri kalanı şu şekilde düzenlenmiştir. Bölüm II' de ilgili literatüre yer verilmiş ve Bölüm III' de kullanılan veri seti anlatılmıştır. Bölüm IV dahilinde uygulanan yöntem detayları açıklanmıştır. Bölüm V' de ise elde edilen sonuçlar paylaşılmıştır. Son olarak Bölüm VI' da çalışma kısaca özetlenmiş ve gelecek çalışmalara yönelik bazı fikirlerden bahsedilmiştir.

II. ARKAPLAN VE ILGILI LITERATÜR

Biyomedikal bilgi keşfi, sağlık sektöründeki bilimsel keşiflerin hızını arttırıcı bir rol üstlenmekte ve hastalıkların anlaşılması ve tedavisi üzerinde önemli etkiler oluşturmaktadır. Bu nedenle, araştırmacılar bu alandaki gelişmeleri yakından takip etmekte ve yeni bilgi keşfi yöntemlerini geliştirmek için çaba sarf etmektedirler. Teknik olarak biyomedikal bilgi keşfi, biyoloji ve tıp gibi bilim alanlarından gelen büyük ve karmaşık veri setleri üzerinde bilgi çıkarma sürecidir [7]. Bu süreç, genetik verilerden klinik verilere kadar çeşitli biyomedikal kaynaklardan elde edilen bilgilerin entegrasyonunu içerir [8]. Bu alandaki temel amaç, geniş veri setleri arasındaki gizli ilişkileri ve desenleri ortaya çıkararak yeni biyomedikal bilgilerin keşfedilmesidir [9,10]. Biyomedikal bilgi keşfi, bir dizi teknik ve metodolojiyi içerir. Örneğin, veri madenciliği yöntemleri, yapay zeka algoritmaları ve ağ analizi teknikleri, büyük veri setlerindeki önemli bilgileri ortaya çıkarmak için kullanılır. İlgili çalışmalar arasında, benzer bilgi keşfi amaçlarını taşıyan birçok araştırma bulunmaktadır. Örneğin, gen ekspresyonu verilerinden hastalık ilişkilerini çıkarmaya yönelik yapılan çalışmalar [10], yeni tedavi yöntemlerinin geliştirilmesine olanak tanımaktadır. Ayrıca, ilaç keşfi [11], hastalık sınıflandırması [12] ve genetik varyasyonların analizi [13, 14] gibi konular üzerinde odaklanan birçok önemli araştırma mevcuttur.

III. VERI SETI DETAYLARI

Bu çalışmada PubMed veri deposunda bulunan medikal alanda yayınlanmış olan çalışmalar (makaleler, konferans bildirileri, vb.) temel veri öğeleri olarak kullanılmıştır. Veri setini türetmek için PubMed dijital deposunda halka açık olan medikal çalışmalar hedef hastalık isimleri anahtar kelime şeklinde belirtilerek aranıp elde edilmiştir. Çalışmaya konu olan hedef hastalıklar "**Tip-2 Diyabet**", "**COVID-19**", "**Obezite**", "**İnflamatuar Bağırsak Hastalığı**" ve "**Kolon Kanseri**" şeklindedir. Özetle, PubMed üzerinde belirlenen hedef hastalıkların literatüründeki güncel bilgilere erişim sağlamak amacıyla Python programlama dili kullanılarak Pymed API aracılığıyla özet ve başlık bilgisi boş olmayan çalışmalar çekilmiştir. Bu süreç, son iki ve üç yıl içerisinde yayınlanmış yayınları içermek üzere sırasıyla 730 gün ve 1095 gün süreleriyle sınırlandırılmıştır. Pymed API' ın tek seferde yapabileceği işlem sınırı 9999 olması sebebiyle, bu işlem her bir hastalık ile ilgili çalışma yayınlanma yoğunluğuna uygun olarak parçalı şekilde gerçekleştirilmiştir.

Elde edilen veriler, her bir hastalık için ayrı ayrı toplamıştır. Örneğin, "**Obezite**" için son üç yıl içinde yayınlanmış 90.485, "**Tip 2 Diyabet**" için son üç yılda yayınlanmış 59.676 makale, "**COVID-19**" için son iki yılda yayınlanmış 113.431 makale, "**Kolon Kanseri**" için son üç yılda yayınlanmış 60.176 makale, yine "**İnflamatuar Bağırsak Hastalığı**" için son üç yılda yayınlanmış 27.303 makale toplanmıştır. Python kodu kullanılarak çekilen özet bilgileri, homojen bir yapı oluşturması ve sonraki adımlarda işlenecek olması nedeniyle tek bir ".json" dosyasında birleştirilmiştir.

IV. METODOLOJI

Bu bölümde yapılan çalışmanın teknik yöntemsel detayları takip eden alt başlıklar halinde anlatılmıştır.

A. Kelime Analizi ve Kelime Filtreleme Adımı

İlk olarak .json formatında birleştirilen özetler içerisindeki anlamsız kelimeler (örn. *the, and, a, an, what, which* vb.) NLTK (Natural Language Toolkit) metin madenciliği kütüphanesi kullanılar filtrelenmiştir. Sonrasında anlamlı kelimeleri belirlemek üzere, en çok tekrar eden ve en az üç harften oluşan 500 kelime/terim Python dilinde tarafımızca yazılan bir program yardımı ile filtrelenerek yeni bir .json formatında dosyada kaydedilmiştir. Böylece bağlamı yansıtan tekrar frekansı yüksek anlamlı 500 terim/kelime elde edilmiştir.

B. Terim-CUI (Concept Unique Identifier) Eşleştirme Adımı

Bu aşamada her bir hastalık için belirlenen ve çalışmalarda geçen kelimelerin UMLS sisteminin API' ı ile ilgili konsept eşsiz tanımlaycı eşleştirme süreci gerçekleştirilmiştir. Unified Medical Language System (UMLS), biyomedikal ve sağlık alanındaki çeşitli terminoloji ve kodlama sistemlerini entegre etmek amacıyla geliştirilmiş bir bilgi kaynağıdır [15]. UMLS Metathesaurus, farklı sağlık terminolojileri arasında benzerlikleri ve ilişkileri tanımlayan geniş bir meta-terim koleksiyonunu içermektedir. Bu koleksiyon, tıp literatürü, hastalık sınıflandırmaları, ilaç isimleri, laboratuvar testleri ve diğer sağlık konularını kapsayan çeşitli kaynaklardan alınan terimleri kapsamaktadır. UMLS Metathesaurus, bu heterojen terminoloji sistemlerini birleştirerek, farklı sağlık bilgi kaynakları arasında etkileşimi kolaylaştırmakta ve her terimin benzersiz bir "CUI" (Concept Unique Identifier) ile tanımlandığı bir sistem sunmaktadır. Bu sayede farklı terminolojilerdeki benzer kavramları eşleştirmeye olanak tanımaktadır. Her bir hastalık için belirlenen önemli 500 kelimenin/terimin CUI eşleştirilmesi amacıyla, UMLS API kullanılmıştır. Bu süreç, varlıklar arası ilişkilerin ortaya çıkarılması için son derece kritik olduğu için her bir varlığa ait benzersiz CUI' lerin başarıyla elde edilmesiyle sağlanmıştır.

C. SemMedDB Veri Tabanı ve İlişki Filtreleme Adımı

SemMedDB, biyomedikal çalışmalardan (başlıklar ve özetler kullanılarak) elde edilen (*özne, yüklem/ilişki, nesne*) ilişkilerinin tutulduğu geniş bir veritabanıdır [16]. Amerika' da bulunan Ulusal Tıp Kütüphanesi (National Library of Medicine - NLM) tarafından kullanıma sunulan halka açık bir kaynak olup, biyomedikal metinden "anlamsal tahminleri" çıkarmak için kural tabanlı bir NLP aracı olan SemRep' i kullanmaktadır [17]. Bu veri tabanı, PubMed arama sistemi aracılığıyla sağlanan tüm biyomedikal çalışmalarda SemRep aracı çalıştırılarak üretilmiş ve bu şekilde elde edilen biyomedikal ilişkilere semantik tripletler denmektedir. Bu ilişkilerin elde edilmesi SemRep aracının, varlıkların (özne ve nesnelerin) metinsel olarak belirtilmesini tekil UMLS Metathesaurus kavramlarına normalleştirerek ve yüklemleri/ilişkileri de

UMLS anlam ağında mevcut olanlara dayandırarak gerçekleştirmektedir. SemMedDB' de bulunan kayıtlar birden fazla akademik çalışmadan çıkarılabileceği için mükerrer sayıda bulunması durumu sözkonusudur. Mükerrer kayıtların (birden fazla çalışmada bulunmuş olması nedeniyle) bertaraf edilmesi ve veri setinin düzenlenmesi amacıyla, ön işleme ve temizleme adımı gerçekleştirilmiştir. Bu sayede her bir kayıt ayrı birim olarak korunmuş ve mükerrer kayıt sayıları frekans değeri olarak ayrıca tutulmak üzere filtreleme yapılmıştır. Bir sonraki filtreleme aşaması olarak SemmedDB üzerinde bulunan 60'ın üzerindeki eşsiz ilişki türünden Tablo I' de belirtilen 30 ilişki türü seçilerek bu ilişkileri içerecek şekilde bir .csv dosyası oluşturulmuştur. Bu sayede kullanılması planlanan ilişki türleri ile yanıltıcı olabilecek gürültü problemine neden olabilecek kayıtlar filtrelenmiştir.

TABLO I. Çizge Yapısı için Belirlenen İlişkiler

PROCESS_OF	COEXISTS_WITH	PREVENTS
ISA	DIAGNOSES	INHIBITS
CAUSES	PREDISPOSES	higher_than
LOCATION_OF	STIMULATES	PART_OF
METHOD_OF	CONVERTS_TO	lower_than
PRODUCES	ASSOCIATED_WITH	DISRUPTS
COMPLICATES	OCCURS_IN	MEASURES
AFFECTS	ADMINISTERED_TO	AUGMENTS
TREATS	MEASUREMENT_OF	PRECEDES
USES	INTERACTS_WITH	same_as

D. SemMedDB ile Bilgi Çizgesi Oluşturma ve Sorgulama

Önceki adımda bahsedilen 30 ilişki türü üzerinden filtrelenmiş olan SemMedDB kayıtları performans gerekçesi ve doğruluğu daha yüksek sonuçlar elde etmek amacıyla ilişkilerin tekrar sayılarına göre en az 3, 5, 10 ve 20 kez bilimsel çalışmalardan elde edilme frekans eşik değerleri ile tekrar filtrelenmiştir. Her bir filtre eşik değerine göre farklı yoğunlukta csv formatında nihai veri setleri oluşturulmuştur. Deneysel analiz çalışmaları amacıyla elde edilen csv dosyalarından frekans değerleri 3 ve 20 olan biyomedikal bilgi çizge yapıları Neo4j platformuyla oluşturulmuştur. Sonrasında Neo4J' e özgü bir sorgulama dili olan Cypher ile yazılan sorgular yardımıyla belirlenen hastalıklardan "Tip 2 Diyabet" (ilgili CUI bilgisi '**C0011860**' olan) için sorgular oluşturulmuş ve çalıştırılmıştır. Her bir ilişkinin anlamsal durumuna göre hedef hastalık CUI bilgisi özne veya nesne pozisyonunda olabilmektedir. Örneğin, *TREATS* ilişkisi ele alındığında hastalık CUI bilgisinin sadece nesne pozisyonunda yer alması anlamlı olmaktadır. Aynı şekilde *COEXISTS_WITH* ilişki türü için hedef hastalık CUI bilgisi her iki pozisyonda da yer alabilmektedir. Bu anlamda "Tip 2 Diyabet" için frekans değeri 3 olan veri seti ile oluşturulan çizge yapısı üzerinde çalıştırılan bazı sorgular aşağıdaki gibidir.

1. **MATCH** (subject:**Node**) - [r:**RELATION** {type: 'AFFECTS'}] -> (object:**Node** {id: '**C0011860**'}) **RETURN** subject, r, object;

2. **MATCH** (subject:**Node** {id:'**C0011860**'}) - [r:**RELATION** {type: ASSOCIATED_WITH}] -> (object:**Node**) **RETURN** subject, r, object;

3. **MATCH** (subject:**Node**) - [r:**RELATION** {type: TREATS}]-> (object:**Node** {id: '**C0011860**'}) **RETURN** subject, r, object;

4. **MATCH** (subject:**Node**) - [r:**RELATION** {type: CAUSES}] -> (object:**Node** {id: '**C0011860**'}) **RETURN** subject, r, object;

5. **MATCH** (subject:**Node**) - [r:**RELATION** {type: STIMULATES}]-> (object:**Node** {id: '**C0011860**'}) **RETURN** subject, r, object;

6. **MATCH** (subject:**Node** {id:'**C0011860**'}) - [r:**RELATION** {type: COEXISTS_WITH}] -> (object:**Node**) **RETURN** subject, r, object;

V. SONUÇLAR VE DEĞERLENDIRME

Bu kısa araştırma bildirisinde deneysel çalışma amacıyla önceden belirlenen hedef hastalıklardan Tip 2 Diyabet hastalığı için gerçekleştirilen akademik çalışmalardan çizge yapısına dönüşüm ve çizge yapısı üzerinden sorgularla hedef hastalık için anlamlı ilişkiler bulma süreci anlatılmıştır. Bir önceki bölümde raporlanan hedef hastalık Cypher sorguları sonucu bulunan ilişkiler Tablo II' de gösterilmiştir. Örneğin, bir numaralı sorgu sonucuna göre Tablo II' deki ilk üç kayıt bulunan sonuçlardır. Birinci kayıta göre TCF7L2 geni Tip 2 Diyabet hastalığını etkilemektedir. Bu sonucu PubMed de yaptığımız bir arama ile teyit ettiğimizde ilk sırada "***The Role of TCF7L2 in Type 2 Diabetes***" isimli 2020 yılında basılan makale çıkmaktadır. Bir diğer örnek olarak altı numaralı sorgu sonucu olarak (*hedef hastalık özne pozisyonunda bulunmakta*) iki farklı *COEXISTS_WITH* ilişkisi bulunmuştur. Detaylı kontrol ettiğimizde Hyperglycemia (*yüksek kan şekeri*) ve Senile cataract (*yaşlılık kataraktı*) hastalıkları hedef hastalık Tip 2 Diyabet ile beraber bulunan hastalıklardır. Her iki sonucu da destekleyen PubMed' de kayıtlı sırasıyla "***The current role of sodium-glucose cotransporter 2 inhibitors in type 2 diabetes mellitus management***" ve ayrıca"***Cataract in diabetes mellitus***" makaleleri bulunmaktadır. Benzer şekilde diğer sorgular sonucunda bulunan sonuçlar da PubMed üzerinde bulunan yayınlarla onaylanabilmektedir. Bu sonuçlar özetle önerilen metodolojinin doğruluğunu göstermektedir.

TABLO II. Sorgu sonucu alınan Tip 2 Diyabet ilişkileri

Özne (CUI - İsim[a])	**İlişki**	**Nesne (CUI - İsim[a])**	**Yayın Sayısı**
C1420644 - TCF7L2 gene	*AFFECTS*	**C0011860**	15
C1456408 - liraglutide	*AFFECTS*	**C0011860**	7
C0021753 - interleukin-1, beta	*AFFECTS*	**C0011860**	4
C1416612 - KCNQ1 gene	*ASSOCIATED_WITH*	**C0011860**	13
C1537896 - MIR375 gene	*ASSOCIATED_WITH*	**C0011860**	5
C0079488 - Helicobacter pylori	*ASSOCIATED_WITH*	**C0011860**	5
C3272640 - Anagliptin	*TREATS*	**C0011860**	8
C2917359 - GLP-1 Receptor Agonist	*TREATS*	**C0011860**	369
C0056331 - cordycepin	*TREATS*	**C0011860**	5
C0011860	*COEXISTS_WITH*	C0020456, Hyperglycemia	78
C0011860	*COEXISTS_WITH*	C0036646, Senile cataract	4

[a]C0011860, Tip 2 Diyabet hastalığını tanımlamaktadır.

VI. ÖZET VE GELECEK ÇALIŞMALAR

Bu çalışmada literatür tabanlı bilgi erişim yöntemlerinden metin-çizge dönüşümü işlemi üzerinden çizge sorgulama süreçleriyle hedef hastalıklar için anlamlı ilişki çıkarımı yapılmıştır. Özellikle hedef hastalık grubundan Tip 2 Diyabet için deneysel çalışma sonuçları raporlanmıştır. Elde edilen ilişkilerin doğruluğu gerçekleştirilen yöntemin başarılı olduğunu ispatlamıştır. Ayrıca önerilen yöntem, her bir hastalık için manuel yordamla PuMed web arayüzü ile yapılacak olan sorgulardan çok daha etkin şekilde ilişkileri bulmaktadır. Gelecekte ilgili çalışma alanları olarak aşağıdaki hedefler planlanmaktadır.

- Tekil ikili bağlantılardan oluşan sorguların genişletilerek daha sofistike sorgularla olası bilinmeyen yeni ilişkilerin keşfi yapılabilecektir.
- UMLS semantik ağındaki her ilişki bir dizi alan semantik tip kısıtlamasına sahiptir. NLM uzman görüşüne dayanarak her bir ilişki için hangi tür varlıkların özne ve nesne rolünü üstlenebileceğini belirlenmiştir. Bu anlamda *TREATS* ilişki sonuçlarına göre semantik tipleri bulunan CUI' ler ile yeni aday ilaçlar klinik çalışma öncesi bulunabilecektir.
- PubMed API üzerinden çok daha özelleşmiş veri çekme yöntemleri ile ilgili spesifik terim veya terimleri içinde barındıran, belli bir yıla kadar yayınlanmış çalışmaları içerecek şekilde yayınlar çekilerek önerilen yöntem tekrar denenebilecektir.
- Benzer analizler diğer hastalıklar için de yapılacaktır.

KAYNAKLAR

[1] Gantz, J. and Reinsel, D., 2012. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012), pp.1-16.

[2] Vijayan, V., Connolly, J.P., Condell, J., McKelvey, N. and Gardiner, P., 2021. Review of wearable devices and data collection considerations for connected health. Sensors, 21(16), p.5589.

[3] Lv, Z. and Qiao, L., 2020. Analysis of healthcare big data. Future Generation Computer Systems, 109, pp.103-110.

[4] Kolukisa, B., Dedeturk, B.K., Dedeturk, B.A., Gulsen, A. and Bakal, G., 2021, September. A Comparative Analysis on Medical Article Classification Using Text Mining & Machine Learning Algorithms. In 2021 6th International Conference on Computer Science and Engineering (UBMK) (pp. 360-365). IEEE.

[5] Lu, Z., 2011. PubMed and beyond: a survey of web tools for searching biomedical literature. Database, 2011, p.baq036.

[6] Guia, J., Soares, V.G. and Bernardino, J., 2017, April. Graph Databases: Neo4j Analysis. In ICEIS (1) (pp. 351-356).

[7] Rather, N.N., Patel, C.O. and Khan, S.A., 2017. Using deep learning towards biomedical knowledge discovery. Int. J. Math. Sci. Comput.(IJMSC), 3(2), pp.1-10.

[8] Yousef, M., Kumar, A. and Bakir-Gungor, B., 2020. Application of biological domain knowledge based feature selection on gene expression data. Entropy, 23(1), p.2.

[9] Mukherjea, S., 2005. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Briefings in bioinformatics, 6(3), pp.252-262.

[10] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P. and Kelley, D.R., 2021. Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods, 18(10), pp.1196-1203.

[11] Bakal, G., Kilicoglu, H. and Kavuluru, R., 2019. Non-negative matrix factorization for drug repositioning: experiments with the repoDB dataset. In AMIA Annual Symposium Proceedings (Vol. 2019, p. 238). American Medical Informatics Association.

[12] Bakal, G. and Kavuluru, R., 2015, December. Predicting treatment relations with semantic patterns over biomedical knowledge graphs. In International Conference on Mining Intelligence and Knowledge Exploration (pp. 586-596). Cham: Springer International Publishing.

[13] Peng, C., Zhang, H., Ren, J., Chen, H., Du, Z., Zhao, T., Mao, A., Xu, R., Lu, Y., Wang, H. and Chen, X., 2022. Analysis of rare thalassemia genetic variants based on third-generation sequencing. Scientific Reports, 12(1), p.9907.

[14] Yousef, M., Sayıcı, A. and Bakir-Gungor, B., 2021. Integrating gene ontology based grouping and ranking into the machine learning algorithm for gene expression data analysis. In Database and Expert Systems Applications-DEXA 2021 Workshops: BIOKDD, IWCFS, MLKgraphs, AI-CARES, ProTime, AISys 2021, Virtual Event, September 27–30, 2021, Proceedings 32 (pp. 205-214). Springer International Publishing.

[15] Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), pp.D267-D270.

[16] Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G. and Rindflesch, T.C., 2012. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics, 28(23), pp.3158-3160.

[17] Kilicoglu, H., Rosemblat, G., Fiszman, M. and Shin, D., 2020. Broad-coverage biomedical relation extraction with SemRep. BMC bioinformatics, 21, pp.1-28.