

Araştırma Makalesi**Kimyasal Bileşiklerin Tat Özelliklerinin GNN Tabanlı Modellerle Sınıflandırılması****Ulvi ISGANDARLI*¹, Eyyüp GÜLBANDILAR²**¹Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü, Eskişehir,ORCID No: <https://orcid.org/0009-0002-0046-5775>²Eskişehir Osmangazi Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir,ORCID No: <https://orcid.org/0000-0001-5559-5281>**Anahtar Kelimeler:**Kimyasal Bileşikler,
Tat Sınıflandırması,
Grafik Sinir Ağları,
Grafik Evrişimli Ağlar,
Grafik Dikkat Ağları

Özet: Bu çalışma, kimyasal bileşiklerin tat özelliklerinin (acı veya tatlı) moleküler yapılarından tahmin edilmesini ele almaktadır. Bu amaçla, bileşiklerin kanonik SMILES gösterimlerinden moleküler grafikler oluşturulmuş ve bu grafikler üzerinde GraphSAGE, Graph Convolutional Network (GCN) ve Graph Attention Network (GAT) olmak üzere üç farklı Graph Neural Network (GNN) mimarisi uygulanmıştır. Modeller, Kaggle platformundan elde edilen bir veri seti üzerinde eğitilmiş ve doğruluk, kesinlik, duyarlılık, özgüllük ve F1 skoru gibi metrikler kullanılarak değerlendirilmiştir. Sonuçlar, GNN tabanlı yaklaşımların tat sınıflandırma görevinde etkili olduğunu ve özellikle GraphSAGE modelinin %90.09 doğruluk ve 0.8772 F1 skoru ile en iyi performansı sergilediğini ortaya koymuştur. Bu bulgular, GNN tabanlı modellerin özellikle tat ile ilişkili bileşiklerin hızlı taranması ve değerlendirilmesi gibi görevlerde ilaç keşfi ve gıda bilimi gibi alanlara katkı sağlayabileceğini göstermektedir.

Research Article**Classification of Taste Properties of Chemical Compounds with GNN-Based Models****Keywords:**Chemical Compounds,
Taste Classification,
Graph Neural Networks,
Graph Convolutional Networks,
Graph Attention Networks

Abstract: This study focuses on predicting the taste properties (bitter or sweet) of chemical compounds from their molecular structures. To achieve this, molecular graphs were generated from the canonical SMILES representations of the compounds, and three different Graph Neural Network (GNN) architectures GraphSAGE, Graph Convolutional Network (GCN), and Graph Attention Network (GAT) were applied to these graphs. The models were trained on a dataset obtained from the Kaggle platform and evaluated using metrics such as accuracy, precision, sensitivity, specificity and F1 score. The results demonstrated that GNN-based approaches are effective in taste classification tasks, with the GraphSAGE model showing the best performance, achieving an accuracy of 90.09% and an F1 score of 0.8772. These findings suggest that GNN-based models can contribute to fields such as drug discovery and food science, especially in tasks like the rapid screening and evaluation of compounds related to taste.

1. GİRİŞ

Kimyasal bileşiklerin tat özellikleri, hem sağlık hem de endüstri açısından büyük önem taşımaktadır. Tat ve koku duyuları, kimyasal maddelerin algılanmasını ve ayırt edilmesini mümkün kılan başlıca duyuşal mekanizmalardır [1]. Örneğin, gıda endüstrisi daha lezzetli ürünler üretmek, ilaç endüstrisi ise ilaçların tadını daha kabul edilebilir hale getirmek için çalışmaktadır [2]. Ancak, bir bileşiğin tadını anlamak amacıyla yapılan geleneksel testler genellikle

pahalı, zaman alıcı ve bazen kişiden kişiye değişebilmektedir [3]. Bu nedenle, bir bileşiğin moleküler yapısına bakarak tadını tahmin edebilen daha hızlı ve düşük maliyetli yöntemlere ihtiyaç duyulmaktadır.

Bu bağlamda, moleküler yapıların doğrusal bir karakter dizisi olarak temsilini sağlayan Basitleştirilmiş Moleküler Giriş Satırı Sistemi (SMILES), kimyasal bilgi işlem ve kemoinformatik veri setlerinde yaygın olarak kullanılan bir standarttır ve bu çalışmada temel veri girdisini oluşturmaktadır [4]. Moleküller için standart bir temsil

sağlamak amacıyla Kanonik SMILES dizileri kullanılmaktadır [5]. Grafik Sinir Ağları (GNN), bir molekülün yapısını ve atomlar arasındaki bağlantıları doğru bir şekilde öğrenebilmesi için her bir Kanonik SMILES kodu, atomları düğüm (node) ve kimyasal bağları kenar (edge) olarak tanımlayan bir moleküler grafiğe dönüştürülmelidir. Bu şekilde oluşturulan moleküler grafikler üzerinde çalışan GNN'ler, molekül yapısındaki önemli desenler ile özellikler arasındaki karmaşık ilişkileri öğrenme kapasitesine sahiptir [6].

Literatürde moleküler tat tahmini için çeşitli yöntemler kullanılmıştır. Renxiu Song ve arkadaşları [7] tarafından yürütülen çalışmada, LLM'ler, özellikle GPT-3.5 ve GPT-4'ün, kimyasal bileşiklerin SMILES dizilerinden acı veya tatlı tat özelliklerini tahmin etme yeteneği araştırılmıştır. Elde edilen sonuçlar, bilhassa yapısal benzerliğe dayalı (scaffold) veri ayırma stratejisi uygulandığında, GPT-4 modelinin %86 gibi dikkate değer bir doğruluk oranına ulaşarak tat tahmini görevinde GPT-3.5'e kıyasla üstün ve umut verici bir performans sergilediğini göstermiştir. Song ve arkadaşları [8] 2601 molekül üzerinde gerçekleştirdikleri karşılaştırmalı analizde farklı moleküler temsil tekniklerini değerlendirerek grafik sinir ağı (GNN) tabanlı modellerin tat tahmininde en yüksek başarıyı sağladığını göstermiştir. En başarılı sonuçlar, F1 skoru 0.852, AUROC 0.957 ve AUPRC 0.917 ile FP + GNN konsensüs modelinde elde edilmiştir. Bu model, hem moleküler parmak izlerinin küresel kimyasal bilgilerini hem de GNN'nin topolojik yapıyı yakalama yetisini birleştirerek tat tahmin performansını anlamlı ölçüde artırmıştır. Bo ve arkadaşları [9], moleküler yapı-tat ilişkisini analiz ettikleri çalışmalarında, acı ve tatlı bileşiklerin sınıflandırılmasında derin öğrenme tabanlı modeller geliştirmişlerdir. Yapılan deneylerde MLP-Fingerprint modeli, acı/tatlı sınıflandırmasında test verisi üzerinde %90 doğruluk ve 0.95 AUC değeriyle en yüksek performansı göstermiştir. Pallante ve arkadaşları [10], 25 farklı TAS2R acı tat reseptörü için geliştirdikleri Random Forest sınıflandırıcılarını, ligand-reseptör etkileşimlerinden elde edilen docking skorları ve hidrojen bağları gibi özelliklerle eğitmiştir. Elde edilen modellerin doğruluk performansı, AUC değerlerinin 0.81 ile 0.86 arasında değişmesiyle yüksek düzeyde başarılı bulunmuştur. Siva Sathyaseelan D. N. ve V. Ramanathan [11], tat içeren kimyasal bileşiklerin sınıflandırılmasında çok etiketli ve çok sınıflı sınıflandırma yaklaşımlarını karşılaştırmalı olarak incelemişlerdir. Çok sınıflı sınıflandırma sonuçlarına göre, en yüksek başarı Morgan parmak izi ile eğitilen Classifier Chains modelinde elde edilmiş; bu model F1 skoru 0.6665, precision 0.7485 ve recall 0.6007 değerlerine ulaşmıştır. Androutsos ve arkadaşları [12], bitter, tatlı, umami ve diğer tatları aynı anda tahmin edebilen çok hedefli bir makine öğrenmesi modeli olan VirtuousMultiTaste'ı geliştirmiştir. Modelin test setindeki genel doğruluğu %71.76, F1 skoru %74.32 ve AUC değeri 0.87 olarak raporlanmıştır. Umami tat sınıfında ise model %95.99 doğruluk ve 0.98 AUC ile en yüksek performansı göstermiştir.

Bu çalışmada, kimyasal bileşiklerin tat özelliklerinin (acı veya tatlı) tahmin edilmesi amacıyla, moleküllerin SMILES gösterimlerinden grafik yapılar oluşturmak için

RDKit kütüphanesinden yararlanılmıştır. RDKit, kimyasal bilgi işleme süreçlerinde yaygın olarak kullanılan açık kaynaklı bir araç olup, SMILES dizilerinden atom ve bağ düzeyinde moleküler grafik temsilleri elde edilmesine olanak tanımaktadır [13]. Elde edilen moleküler grafik yapılar, GNN modellerine giriş verisi olarak kullanılmıştır.

Bu kapsamda üç farklı GNN mimarisi karşılaştırılmıştır: GraphSAGE, Grafik Dikkat Ağı (Graph Attention Network - GAT) ve Grafik Konvolüsyonel Ağ (Graph Convolutional Network - GCN). Modeller, RDKit ile oluşturulan moleküler grafikler üzerinde eğitilmiş ve değerlendirilmiştir. Deneysel çalışmalar, Kaggle platformundan temin edilen halka açık bir veri seti kullanılarak gerçekleştirilmiştir. Çalışmanın temel amacı, modern GNN mimarilerinin moleküler tat tahmini gibi özgül bir sınıflandırma problemindeki etkinliğini ortaya koymak ve bu mimarilerin performanslarını karşılaştırmalı olarak değerlendirmektir.

2. MATERYAL VE METOT

Bu bölümde, çalışmada kullanılan veri setinin özellikleri, veri ön işleme adımları, moleküler grafiklerin oluşturulma süreci, uygulanan GNN mimarileri, model eğitim stratejileri ve performans değerlendirme metrikleri ayrıntılı olarak açıklanmaktadır.

2.1. Veri Seti

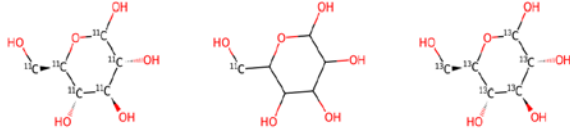
Araştırmada, Kaggle platformunda kamuya açık olarak sunulan "Classify the Bitter or Sweet Taste of Compounds" adlı veri seti kullanılmıştır [14]. Veri seti; bileşik adı, PubChem kimlik numarası (PubChem_CID), moleküler ağırlık, çeşitli 2D ve 3D tanımlayıcılar, SMILES ve Canonical SMILES gösterimleri ile tat özelliğini belirten hedef sınıf (Target) gibi birçok yapısal ve tanımlayıcı alan içermektedir.

Bu çalışmada, tat tahmini doğrudan yapısal temsiller üzerinden gerçekleştirildiğinden analizler yalnızca Canonical SMILES ve Target sütunlarına odaklanmıştır. Diğer moleküler tanımlayıcılar bu çalışma kapsamında dikkate alınmamıştır. Target sütununda "Bitter" (acı) ve "Sweet" (tatlı) olarak etiketlenmiş örnekler değerlendirmeye alınmış, "Non_Bitter_Sweet" olarak işaretlenen örnekler analiz dışında bırakılmıştır. İkili sınıflandırma yaklaşımı doğrultusunda, "Bitter" sınıfı negatif (0), "Sweet" sınıfı ise pozitif (1) olarak kodlanmıştır. Moleküler yapıların standart ve tekil temsili için Canonical SMILES dizileri birincil girdi olarak kullanılmış ve moleküler grafikler türetilmiştir. Uygulanan veri temizleme, filtreleme ve ön işleme adımları sonucunda, toplam 2.120 adet bileşiği içeren nihai bir veri seti elde edilmiştir.

2.2. Moleküler Grafik Temsili

Her bir kanonik SMILES dizisi, RDKit ve PyTorch Geometric [15] kütüphaneleri kullanılarak bir moleküler grafik biçiminde modellenmiştir. Bu süreçte öncelikle RDKit aracılığıyla SMILES dizilerinin geçerliliği kontrol

edilmiş; geçersiz olanlar analiz kapsamı dışında bırakılmıştır. İlk üç bileşiğe ait kanonik SMILES dizilerinin kimyasal yapıları Şekil 1'de görselleştirilmiştir.



Şekil 1. Veri setindeki ilk üç bileşiğin kanonik SMILES gösterimlerine karşılık gelen kimyasal yapıları.

Geçerli SMILES dizileri, otomatik olarak PyTorch Geometric'in Data nesnelerine dönüştürülmüştür. Bu dönüşüm sonucunda, atomlar düğüm (node), kimyasal bağlar ise kenar (edge) olarak temsil edilen grafik yapıları oluşturulmuştur. Her düğüm ve kenar için, modelin işleyebileceği biçimde temel yapısal öznitelikler otomatik olarak çıkarılmıştır. Bu öznitelikler, PyTorch Geometric'in varsayılan dönüştürücüleri tarafından sağlanan atom tipi ve bağ türü gibi temel kimyasal bilgileri içermektedir. Daha karmaşık tanımlayıcıların çıkarımı ise bu çalışmanın kapsamı dışında bırakılmıştır. Sonuç olarak, her molekül grafi; x, edge_index, edge_attr ve y bileşenlerini içeren bir Data nesnesi olarak yapılandırılmıştır.

Son aşamada, oluşturulan grafik veri kümesi model eğitimi için %80 eğitim, %10 doğrulama ve %10 test alt kümelerine rastgele şekilde ayrılmıştır.

2.3. Grafik Sinir Ağı Mimarileri

Bu çalışmada, moleküler tat tahmini görevine yönelik olarak üç farklı GNN mimarisi uygulanmış ve karşılaştırılmıştır: GraphSAGE, GAT ve GCN. Tüm mimariler, Bölüm 2.1'de belirtildiği üzere "Bitter" ve "Sweet" olmak üzere iki sınıfa ayıran bir sınıflandırma görevi için yapılandırılmıştır. Hiperparametreler tüm modeller için sabit tutulmuş; gizli katman boyutu (hidden_channels) 64 ve tam bağlı ara katman boyutu 256 olarak belirlenmiştir.

2.3.1. GraphSAGE Mimarisi

GraphSAGE modeli, her bir düğümün yerel yapısını temsil edebilmek amacıyla, düğümün komşularından örnekleme yaparak bu komşu düğümlerin özelliklerini belirli bir toplama fonksiyonu aracılığıyla birleştiren ve elde edilen temsili merkez düğümün güncellenmesinde kullanan bir grafik sinir ağı mimarisidir [16]. Bu çalışma kapsamında uygulanan GraphSAGE mimarisi, her biri ortalama toplama stratejisi kullanan üç adet SAGEConv katmanından oluşmakta olup; ilk katman, düğüm özelliklerini 128 boyutlu bir temsile dönüştürmekte, ikinci katman bu temsili koruyarak bir artık bağlantı (residual connection) ile desteklemekte, üçüncü katman ise bu temsili 64 boyutuna indirgemektedir. Her SAGEConv katmanından sonra sırasıyla Batch Normalization ve ReLU aktivasyon fonksiyonu uygulanmakta, sonrasında elde

edilen düğüm temsilleri, graf düzeyinde tek bir gömme (embedding) vektörü elde etmek amacıyla ortalama havuzlama (mean pooling) işlemiyle birleştirilmekte; bu vektör sırasıyla 256 boyutlu bir tam bağlı katmana (ardından ReLU aktivasyonu ve %50 dropout) ve iki sınıflı çıktı katmanına yönlendirilmektedir.

2.3.2. Grafik Dikkat Ağı Mimarisi

GAT modeli, komşu düğümlerin özelliklerini toplarken dikkat mekanizmalarını kullanarak her bir komşunun katkısını dinamik biçimde ağırlıklandırmakta ve ayrıca kenar özelliklerini de dikkate alabilmektedir [17]. Bu bağlamda geliştirilen GAT mimarisi üç adet GATConv katmanından oluşmakta; ilk katman, düğüm özelliklerini dört dikkat başlığı (gat_heads = 4) kullanarak ve bu başlıkların çıktısını birleştirerek 256 boyutlu bir temsile dönüştürmekte, ikinci katman bu temsili yine 4 başlıkla koruyarak bir artık bağlantı ile güçlendirmekte, üçüncü katman ise dikkat başlıklarının çıktılarının ortalamasını alarak temsili 64 boyutuna indirgemektedir. Her bir GATConv katmanının ardından Batch Normalization ve ReLU aktivasyon fonksiyonu uygulanmıştır. Düğüm temsilleri, grafik düzeyinde tek bir vektörde özetlenmesi amacıyla ortalama havuzlama yöntemiyle birleştirilmekte ve ardından 256 boyutlu bir tam bağlı katman, ReLU aktivasyonu ve %50 dropout uygulanarak iki sınıflı çıktı katmanına yönlendirilmektedir.

2.3.3. Grafik Konvolüsyonel Ağ

GCN'ler, grafik yapısındaki düğümlerin özelliklerini komşuluk bilgilerini kullanarak güncelleyen katmanlardan oluşur [18]. Model, ortalama komşuluk toplama stratejisi kullanan üç adet NNConv katmanından oluşmakta olup, her bir NNConv katmanı içinde kenar özelliklerini işleyen iki katmanlı bir sinir ağı (nn.Sequential) yer almaktadır. Bu kenar ağı, kenar özelliklerini önce 32 boyutlu bir ara temsile, ardından da NNConv katmanının ağırlık matrisine dönüştürmektedir. İlk NNConv katmanı, giriş düğüm özelliklerini 64 boyutlu bir temsile dönüştürürken; ikinci ve üçüncü katmanlar bu temsili korumakta, her biri artık bağlantılarla (residual connections) desteklenmekte ve sonrasında sırasıyla Batch Normalization ve ReLU aktivasyon fonksiyonları uygulanmaktadır. Elde edilen düğüm temsilleri ortalama havuzlama yöntemi ile grafik seviyesinde özetlenmekte, ardından 256 boyutlu bir tam bağlı katman (ReLU aktivasyonu ve %50 dropout) ve nihai olarak iki sınıflı bir çıktı katmanına aktarılmaktadır.

2.4. Model Eğitimi ve Değerlendirme

Modelin eğitimi PyTorch Lightning kullanılarak yapılmıştır. Model, düğüm özellikleri, kenar özellikleri ve sınıf sayısı gibi parametrelerle yapılandırılmıştır. Doğrulama kaybı en düşük değeri aldığı anda model kaydedilmiş ve eğitim süreci erken durdurma yöntemiyle kontrol edilmiştir. Bu, modelin doğrulama kaybı 20 epoch boyunca iyileşmezse eğitimin durmasını sağlamıştır.

Modelin performansı, test verisi üzerinde çeşitli metriklerle değerlendirilmiştir. Bu metrikler arasında

doğruluk (1), kesinlik (2), hassasiyet (3), özgüllük (4) ve F1 skoru (5) yer almaktadır. Doğruluk, modelin doğru sınıflandırdığı örneklerin oranını gösterirken, kesinlik ve duyarlılık sırasıyla pozitif sınıfların doğruluğunu ve modelin gerçek pozitif sınıflarındaki başarısını yansıtmaktadır [19]. Özgüllük, negatif sınıfların doğru şekilde sınıflandırılma oranını ifade ederken, F1 skoru, modelin kesinlik ve hassasiyet arasında bir denge kurarak genel başarısını gösterir [20]. Yukarıda bahsedilen performans metriklerinin hesaplama formülleri aşağıdaki gibidir:

$$\text{Doğruluk} = \frac{DP+DN}{DP+DN+YP+YN} \quad (1)$$

$$\text{Kesinlik} = \frac{DP}{DP+YP} \quad (2)$$

$$\text{Duyarlılık} = \frac{DP}{DP+YN} \quad (3)$$

$$\text{Özgüllük} = \frac{DN}{DN+YP} \quad (4)$$

$$\text{F1 Skoru} = 2 \times \frac{\text{kesinlik} + \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}} \quad (5)$$

3. BULGULAR

Bu bölümde, GCN, GAT ve GraphSAGE modellerinin tat tahmini görevindeki performanslarına ilişkin sonuçlar sunulmaktadır. Modeller, test seti üzerinde çeşitli değerlendirme metrikleriyle analiz edilmiş ve sonuçlar Tablo 1'de özetlenmiştir.

Tablo 1. GCN, GAT ve GraphSAGE modellerinin test seti üzerindeki performans karşılaştırması.

Model	Doğruluk	Kesinlik	Duyarlılık	Özgüllük	DP	YP	DN	YN	F1 Skor
GCN	0.8868	0.9114	0.809	0.9431	72	7	116	17	0.8571
GAT	0.8632	0.9054	0.7528	0.9431	67	7	116	22	0.8221
GraphSAGE	0.9009	0.9146	0.8427	0.9431	75	7	116	14	0.8772

Tablo 1'de görüldüğü üzere, tüm modeller tat tahmini görevinde başarılı sonuçlar vermiştir. Dikkat çekici bir nokta, tüm modellerin özgüllük (specificity) değerinin 0.9431 ile aynı olmasıdır. Bu durum, her üç modelin de negatif sınıfı (örneğin, tatlı olmayan bileşikler) doğru bir şekilde tanımlamada benzer ve yüksek bir yetkinliğe sahip olduğunu göstermektedir. Ayrıca, tüm modellerin Yanlış Pozitif (FP) sayısının 7 olması, pozitif olarak yanlış sınıflandırdıkları örnek sayısının eşit olduğunu ortaya koymaktadır.

Modellerin bireysel performansları ve karışıklık matrisi sonuçları incelendiğinde, GraphSAGE modelinin en yüksek başarıyı sergilediği gözlemlenmektedir. Bu model, test seti üzerinde 75 Doğru Pozitif (TP), 7 Yanlış Pozitif (FP), 116 Doğru Negatif (TN) ve 14 Yanlış Negatif (FN) değerleri elde etmiştir. Bu dağılım, %90.09 doğruluk, %91.46 kesinlik, %84.27 duyarlılık, %94.31 özgüllük ve 0.8772 F1 skoru gibi üstün metrik sonuçlarına yansımıştır. GraphSAGE'in en yüksek TP ve en düşük FN sayılarına sahip olması, pozitif sınıfı (örneğin, tatlı bileşikler) yakalamada ve genel sınıflandırma performansında diğer modellere göre daha etkili olduğunu göstermektedir.

GCN modeli, GraphSAGE'in ardından ikinci en iyi performansı sergilemiştir. Bu model, 72 TP, 7 FP, 116 TN

ve 17 FN sonucuyla %88.68 doğruluk, %91.14 kesinlik, %80.90 duyarlılık, %94.31 özgüllük ve 0.8571 F1 skoruna ulaşmıştır. GCN modelinin kesinlik değeri GraphSAGE modeline oldukça yakın bir seviyededir; ancak Yanlış Negatif sayısının GraphSAGE'e göre biraz daha yüksek olması, duyarlılık değerinin bir miktar daha düşük kalmasına neden olmuştur.

Değerlendirilen modeller arasında GAT modeli, diğer iki modele kıyasla görece daha düşük bir performans göstermiştir. GAT, 67 TP, 7 FP, 116 TN ve 22 FN değerleriyle %86.32 doğruluk, %90.54 kesinlik, %75.28 duyarlılık, %94.31 özgüllük ve 0.8221 F1 skoru elde etmiştir. Özellikle, GAT modelinin diğerlerine kıyasla daha düşük Doğru Pozitif (67) ve daha yüksek Yanlış Negatif (22) sayısına sahip olması, duyarlılık metriğindeki (%75.28) belirgin düşüklüğü açıklamaktadır.

Bu bulgular, modellerin tat tahmini problemindeki göreceli Etkinliklerini ve sınıflandırma hatalarının dağılımını sayısal olarak ortaya koymakta olup, bu sonuçların nedenleri ve olası çıkarımlar 4. bölümünde daha detaylı bir şekilde ele alınmaktadır.

4. TARTIŞMA VE SONUÇ

Bu çalışma, kimyasal bileşiklerin tat özelliklerini moleküllerin SMILES gösterimlerinden türetilen ve atomları düğüm, kimyasal bağları ise kenar olarak temsil eden grafik yapılar üzerinden sınıflandırmayı hedeflemiştir. GCN, GAT ve

GraphSAGE olmak üzere üç farklı GNN mimarisi kullanılmış ve her biri farklı yapıları ve öğrenme yöntemleri nedeniyle farklı sonuçlar vermiştir. Çalışmamız ile moleküler tat tahmini üzerine yapılan çalışmaların karşılaştırılması Tablo 2’de verilmiştir.

Tablo 2. Moleküler tat tahmini üzerine yapılan çalışmaların karşılaştırılması.

Çalışma	Kullanılan Model/Yöntem	Moleküler Temsil	Tahmin Görevi/Hedef	Performans Metrikleri
[7]	Büyük Dil Modelleri (GPT-4)	SMILES	Acı/tatlı sınıflandırması	GPT-4 doğruluk: %86 (scaffold ayrımı ile)
[8]	FP + GNN Konsensüs Modeli	Moleküler parmak izi	Tat sınıflandırması	F1: 0.852, AUROC: 0.957, AUPRC: 0.917
[9]	Derin öğrenme	Moleküler parmak izi	Acı/tatlı sınıflandırması	Doğruluk: %90, AUC: 0.95
[10]	Random Forest	Docking skorları + hidrojen bağları	TAS2R reseptörleri üzerinden ligand-tat etkileşimleri	AUC: 0.81–0.86
[11]	Classifier Chains	Morgan parmak izi	Tat içeren bileşiklerin çoklu sınıflandırması	F1: 0.6665, Precision: 0.7485, Recall: 0.6007
[12]	VirtuousMultiTaste (çok hedefli makine öğrenmesi modeli)	SMILES	Bitter, tatlı, umami vb. tatların çok hedefli tahmini	Genel doğruluk: %71.76, F1: %74.32, AUC: 0.87, Umami doğruluk: %95.99, AUC: 0.98
Bizim Çalışmamız	GraphSage, GAT, GCN	SMILES	Acı/tatlı sınıflandırması	GCN: doğruluk: 0.8868, GAT: doğruluk: 0.8632, GraphSAGE: doğruluk: 0.9009

GraphSAGE, bu çalışma kapsamında en yüksek başarıyı gösteren model olmuştur. Modelin başarısının temelinde, düğüm komşularından örnekleme yoluyla özellik aktarımı gerçekleştiren benzersiz öğrenme stratejisi yatmaktadır. Bu yaklaşım, her düğüm için yalnızca belirli sayıda komşuyu dikkate alarak bilgi topladığından, gürültüye açık ya da yüksek bağlantılı grafiklerde dahi öğrenme sürecinin bozulmadan ilerlemesini sağlamaktadır [21]. Böylece model, moleküler yapılar arasındaki karmaşık ilişkileri daha etkili bir şekilde temsil ederek hem genel doğrulukta (%90.09) hem de dengeleyici bir performans göstergesi olan F1 skorunda (0.8772) diğer modellere kıyasla daha iyi sonuçlar üretmiştir. Ayrıca modelin yüksek duyarlılık (recall: 0.843) ve kesinlik (precision: 0.915) değerleri, tatlı sınıfa ait bileşikler hem doğru hem de güvenilir biçimde sınıflandırabildiğini göstermektedir. Bu yüksek performans, özellikle moleküler tat tahmini gibi biyokimyasal sınıflandırma problemlerinde GraphSAGE mimarisinin güçlü bir alternatif olduğunu ortaya koymaktadır. Nitekim

literatürde, GraphSAGE mimarisinin; biyolojik ağ analizi [16], protein-protein etkileşimleri ve çok ilişkili biyomedikal ağların modellenmesi [22] ile moleküler özellik tahmini [23] gibi yapı temelli görevlerde başarılı performans sergilediği çeşitli çalışmalarla ortaya konmuştur.

GCN, istikrarlı ve dengeli bir performans sunarak %88.68 doğruluk ve 0.8571 F1 skoru ile ikinci sırada yer almıştır.

Modelin özellikle negatif sınıfları (acı bileşikler) ayırt etmedeki başarısı, yüksek özgüllük ve kesinlik değerleriyle desteklenmektedir. Bu yönüyle GCN, yanlış pozitiflerin minimize edilmesinin önemli olduğu durumlarda uygun bir tercih olabilir.

GAT, düğümler arası ilişkileri dikkat mekanizması aracılığıyla ağırlıklı olarak modelleyebilme yeteneğine sahip olmasına rağmen, bu çalışmada diğer modellere kıyasla nispeten daha düşük bir performans sergilemiştir.

%86.32 doğruluk ve 0.8221 F1 skoru elde eden model, özellikle duyarlılık (0.753) açısından sınırlı kalmıştır. Bu durum, dikkat mekanizmasının kullanılan moleküler temsil biçimi (SMILES'den türetilen grafik yapılar) ile olan etkileşiminin yeterince optimize edilmemiş olabileceğini düşündürmektedir. Ancak, literatürde dikkat mekanizmalarının özellikle daha büyük ve çeşitli veri kümelerinde etkili sonuçlar verdiği bildirilmektedir. Bu nedenle, GAT mimarisinin potansiyelinin mevcut veri kümesinin ötesinde, daha karmaşık ve heterojen moleküler yapılar içeren senaryolarda daha belirgin şekilde ortaya çıkabileceği değerlendirilmektedir. İlerleyen çalışmalarda hiperparametre optimizasyonu, veri artırma stratejileri veya dikkat modülünün mimari düzeyde geliştirilmesi yoluyla bu modelin performansı artırılabilir.

Genel olarak, elde edilen bulgular GNN tabanlı modellerin, moleküllerin temel yapısal temsillerinden yola çıkarak dahi, tat sınıflandırması gibi karmaşık görevlerde güçlü bir yaklaşım sunduğunu ortaya koymaktadır. Özellikle GraphSAGE modelinin gösterdiği üstün başarı, bu mimarinin ilaç keşfi, toksikoloji ve gıda kimyası gibi alanlarda uygulanabilirliğini desteklemektedir.

Gelecek çalışmalarda, daha büyük ve çeşitli veri kümelerinin kullanımı, moleküler özelliklerin (örneğin atom türü, bağ türü, polarite vb.) daha ayrıntılı şekilde modellenmesi ve gelişmiş GNN mimarilerinin (örneğin Graph Transformer, DimeNet, MPNN) değerlendirilmesi, performansı daha da ileriye taşıyabilir.

Etik Hususlar

Etik kurallara uyum

Bu araştırma, bilimsel etik kurallar çerçevesinde gerçekleştirilmiş olup; çalışma sürecinin tüm aşamalarında akademik dürüstlük ilkelerine uygun hareket edilmiştir. Çalışmada kullanılan veri, halka açık bir veri seti [14] üzerinde gerçekleştirilmiş olup, bu veri seti üzerinden yapılan analizler doğrultusunda elde edilen sonuçlar raporlanmıştır.

Çıkar çatışması

Yazarlar, bu çalışma kapsamında herhangi bir çıkar çatışmasının bulunmadığını açıkça beyan etmektedir.

KAYNAKÇA

- [1] Schieberle, P., & Hofmann, T. (2011). Mapping the combinatorial code of food flavors by means of molecular sensory science approach. In *Chemical, Sensory and Technological Properties* (pp. 413-438). Informa UK Limited.
- [2] Schwartz, C., Chabanet, C., Lange, C., Issanchou, S., & Nicklaus, S. (2011). The role of taste in food acceptance at the beginning of complementary feeding. *Physiology & behavior*, 104(4), 646-652.
- [3] Mennella, J. A., Spector, A. C., Reed, D. R., & Coldwell, S. E. (2013). The bad taste of medicines: overview of basic research on bitter taste. *Clinical therapeutics*, 35(8), 1225-1246.
- [4] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1), 31-36.
- [5] O'Boyle, N. M. (2012). Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics*, 4, 1-14.
- [6] Wang, Y., Li, Z., & Barati Farimani, A. (2023). Graph neural networks for molecules. In *Machine learning in molecular sciences* (pp. 21-66). Cham: Springer International Publishing.
- [7] Song, R., Liu, K., He, Q., He, F., & Han, W. (2024). Exploring bitter and sweet: the application of large language models in molecular taste prediction. *Journal of Chemical Information and Modeling*, 64(10), 4102-4111.
- [8] Song, Y., Chang, S., Tian, J., Pan, W., Feng, L., & Ji, H. (2023). A comprehensive comparative analysis of deep learning based feature representations for molecular taste prediction. *Foods*, 12(18), 3386.
- [9] Bo, W., Qin, D., Zheng, X., Wang, Y., Ding, B., Li, Y., & Liang, G. (2022). Prediction of bitterant and sweetener using structure-taste relationship models based on an artificial neural network. *Food Research International*, 153, 110974.
- [10] Pallante, L., Cannariato, M., Vezzulli, F., Malavolta, M., Lambri, M., & Deriu, M. A. (2023). Machine learning aided molecular modelling of taste to identify food fingerprints. *Chemical Engineering Transactions*, 102(2023), 283-288.
- [11] Ramanathan, V., & DN, S. S. (2025). Predicting Molecular Taste: Multi-Label and Multi-Class Classification. *bioRxiv*, 2025-05.
- [12] Androutsos, L., Pallante, L., Bompotas, A., Stojceski, F., Grasso, G., Piga, D., ... & Mavroudi, S. (2024). Predicting multiple taste sensations with a multiobjective machine learning method. *npj Science of Food*, 8(1), 47.
- [13] Landrum, G. (2013). Rdkit documentation. *Release*, 1(1-79), 4.
- [14] Farmer, R. (2022). *Classify the bitter or sweet taste of compounds* [Veri seti]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/42341_93
- [15] Fey, M., & Lenssen, J. E. (2019). *Fast graph representation learning with PyTorch Geometric*. arXiv. <https://arxiv.org/abs/1903.02428>
- [16] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.
- [17] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph

- attention networks. *arXiv preprint arXiv:1710.10903*.
- [18] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- [19] Yacoub, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).
- [20] Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of clinical psychology*, 44(6), 1013-1023.
- [21] Oh, J., Cho, K., & Bruna, J. (2019). Advancing graphsage with a data-driven node sampling. *arXiv preprint arXiv:1904.12935*.
- [22] Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), i457-i466.
- [23] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2019). Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.