Bölüm 8

Bilgisayar Bilimleri ve Mühendisliğinde Dönüşüm: Üretken Yapay Zekâ, Güvenli Sistemler ve Ölçeklenebilir Hesaplama

Furkan Atlan[1]

Özet

Bu çalışma, 2010-2026 döneminde bilgisayar bilimleri ve mühendisliğinde yaşanan dönüşümü; üretken yapay zekâ, güvenli yapay zekâ sistemleri ve ölçeklenebilir hesaplama altyapıları ekseninde bütüncül biçimde incelemektedir. İlk olarak, veri yoğun dünyada artan hesaplama gereksinimleri, donanım sınırları ve yazılım merkezli optimizasyon yaklaşımları ele alınmakta; sağlık, enerji, finans ve savunma gibi alanlarda ortaya çıkan disiplinler arası etkiler tartışılmaktadır. Ardından üretken yapay zekâ ve temel modeller kapsamında büyük dil modelleri, çok modlu sistemler, Edge AI, hafif model mimarileri ve açıklanabilir yapay zekâ yaklaşımları değerlendirilmekte; model güvenilirliği ve halüsinasyon problemi, çağdaş yapay zekâ sistemlerinin temel sınırlılıkları arasında konumlandırılmaktadır. Bölümün güvenlik odaklı kısmında adversarial saldırılar, veri/model zehirleme, yapay zekâ ile siber güvenlik kesişimi, düzenleyici çerçeveler ve algoritmik adalet konuları ele alınmaktadır. Son olarak GPU/TPU hızlandırmalı sistemler, bulut ve hibrit mimariler, HPC kümeleri, Slurm tabanlı kaynak yönetimi, edge-cloud iş bölümü ve enerji verimli hesaplama başlıkları üzerinden modern hesaplama altyapılarının teknik ve stratejik boyutları tartışılmaktadır. Genel olarak bölüm, yapay zekâ çağında güvenilir, açıklanabilir, sürdürülebilir ve ölçeklenebilir bilişim sistemlerinin geliştirilmesinin bilgisayar bilimlerinin temel yönelimlerinden biri hâline geldiğini ortaya koymaktadır.

1 Dr. Öğr. Üyesi, Burdur Mehmet Akif Ersoy Üniversitesi Yönetim Bilişim Sistemleri Bölümü, fatlan@mehmetakif.edu.tr , https://orcid.org/0000-0003-1602-1941

https://doi.org/10.58830/ozgur.pub1230.c4969

1. Bilgisayar Bilimlerinde Paradigma Değişimi

1.1. 2010-2026 Yılları Arasındaki Teknolojik Kırılmalar

2010-2026 yılları arası dönem, bilgisayar bilimlerinde hem kuramsal yaklaşımların hem de uygulama alanlarının önemli ölçüde yeniden şekillendiği bir zaman dilimi olarak değerlendirilmektedir. Bu süreçte veri üretim hızının artması, hesaplama altyapılarının gelişmesi ve yapay zekâ (AI) algoritmalarındaki ilerlemeler, bilgisayar bilimlerini yalnızca teknik bir disiplin olmaktan çıkararak çok sayıda alanla kesişen disiplinler arası bir araştırma alanına dönüştürmüştür. Özellikle büyük veri (Big Data) ekosisteminin ortaya çıkışı, bulut bilişim altyapılarının yaygınlaşması ve dağıtık sistem mimarilerinin olgunlaşması, yüksek hacimli verilerin işlenmesini mümkün kılmış ve veri odaklı karar verme süreçlerinin bilimsel ve endüstriyel alanlarda yaygınlaşmasına katkı sağlamıştır (Bibri, 2019).

Bu dönemde dikkat çeken en önemli teknolojik kırılmalardan biri AI ve derin öğrenme alanında yaşanmıştır. Özellikle 2012 yılında derin sinir ağlarının görüntü tanıma problemlerinde gösterdiği performans artışı, AI araştırmalarında yeni bir paradigma oluşturmuştur. Takip eden yıllarda konvolüsyonel sinir ağları (CNN), tekrarlayan sinir ağları (RNN) ve transformer tabanlı mimariler gibi modellerin geliştirilmesi; doğal dil işleme, bilgisayarla görme ve konuşma tanıma gibi alanlarda önemli ilerlemelerin gerçekleşmesini sağlamıştır (Chinnaiyan vd., 2025). 2020'li yıllarla birlikte ise büyük ölçekli dil modelleri (Large Language Models-LLM) ve üretken AI (Generative AI) sistemleri, bilgi üretimi, içerik oluşturma ve insan–makine etkileşimi gibi alanlarda yeni araştırma ve uygulama fırsatları ortaya çıkarmıştır (Zhou vd., 2024).

Donanım tarafında yaşanan gelişmeler de bu dönüşümü hızlandıran önemli faktörlerden biri olmuştur. Grafik işlem birimleri (GPU), tensör işlem birimleri (TPU) ve özel amaçlı AI hızlandırıcılarının geliştirilmesi, büyük ölçekli AI modellerinin eğitilmesini mümkün kılmıştır. Bunun yanında yüksek performanslı hesaplama (High Performance Computing - HPC) altyapılarının ve paralel işlem mimarilerinin gelişmesi, bilimsel hesaplamalarda ve veri yoğun uygulamalarda önemli performans kazanımları sağlamıştır (Deelman vd., 2025). Ayrıca uç bilişim (Edge Computing) ve nesnelerin interneti (IoT) teknolojileri sayesinde hesaplama süreçleri merkezi sistemlerden dağıtık ve akıllı cihazlara doğru kaymaya başlamıştır.

2010-2026 döneminde bilgisayar bilimlerindeki dönüşüm yalnızca teknik gelişmelerle sınırlı kalmamış, aynı zamanda güvenlik, etik ve düzenleyici çerçeveler açısından da yeni tartışmaları beraberinde getirmiştir. AI sistemlerinin

karar verme süreçlerindeki rolünün artması; açıklanabilir AI (Explainable AI), algoritmik adalet ve veri güvenliği gibi konuların önemini artırmıştır (Altukhi ve Pradhan, 2025). Bunun yanı sıra siber güvenlik tehditlerinin karmaşıklaşması, kuantum hesaplama araştırmalarının hız kazanması ve sürdürülebilir bilişim yaklaşımlarının gündeme gelmesi, bilgisayar bilimlerinin gelecekteki araştırma yönelimlerini belirleyen temel unsurlar arasında yer almaktadır. Bu bağlamda 2010-2026 yılları arası dönem, bilgisayar bilimlerinde paradigma değişimlerinin yaşandığı ve dijital dönüşümün küresel ölçekte hız kazandığı kritik bir kırılma noktası olarak değerlendirilmektedir.

1.2. Donanım Sınırları ve Yazılım Merkezli Optimizasyon

Bilgisayar bilimlerinde son on beş yılda yaşanan en önemli dönüşümlerden biri, donanım ölçeklenmesinin fiziksel sınırlarına yaklaşılmasıyla birlikte performans artışının giderek yazılım temelli optimizasyonlara dayanır hâle gelmesidir. Moore Yasası'nın yavaşlaması, transistor boyutlarının küçülmesiyle ortaya çıkan enerji tüketimi, ısınma ve güvenilirlik sorunları, işlemci mimarilerinin yalnızca donanım geliştirmeleriyle performans kazanmasını zorlaştırmıştır. Özellikle modern çok çekirdekli sistemlerde artan entegrasyon yoğunluğu, işlemcileri geçici ve kalıcı hatalara karşı daha hassas hâle getirmiştir. Bu durum, sistem doğruluğunu korumak amacıyla mimari düzeyde hata toleransı ve yazılım destekli güvenilirlik mekanizmalarının geliştirilmesini gerekli kılmıştır (Keller, 2025).

Donanım sınırlamalarının belirginleşmesiyle birlikte araştırmacılar, performans ve güvenilirlik hedeflerine ulaşmak için donanım-yazılım ortak tasarımına dayalı optimizasyon stratejilerine yönelmiştir. Çok çekirdekli işlemcilerde paralel yürütme yetenekleri önemli performans avantajları sunsa da bu mimariler, yazılım tarafından etkin şekilde yönetilmediğinde beklenen verimliliği sağlayamamaktadır. Bu nedenle görev zamanlama, hata tespiti, redundant çoklu iş parçacığı (redundant multithreading) ve geri alma (rollback) mekanizmaları gibi yazılım tabanlı teknikler, sistem güvenilirliğini artırmak için yaygın olarak kullanılmaktadır. Yazılım seviyesinde gerçekleştirilen bu tür optimizasyonlar, donanım maliyetini artırmadan hata toleransı sağlayabilmeleri nedeniyle özellikle modern işlemci mimarilerinde önemli bir araştırma alanı hâline gelmiştir (Keller, 2025; De Souza, 2025).

Benzer şekilde, günümüz dijital ürün ve sistem mimarilerinde donanım ile yazılım bileşenlerinin ayrıştırılması (hardware-software decoupling) önemli bir tasarım yaklaşımı olarak ortaya çıkmıştır. Yazılım tanımlı ürünler (Software-Defined Products) yaklaşımında sistem fonksiyonlarının büyük bölümü yazılım katmanında tanımlanmakta ve donanım yalnızca temel altyapıyı

sağlamaktadır (Barwasser vd., 2024). Bu yaklaşım sayesinde ürün özellikleri, donanım değişikliğine ihtiyaç duyulmadan yazılım güncellemeleri aracılığıyla geliştirilebilmekte ve sistemler daha esnek hâle getirilebilmektedir. Özellikle otomotiv, tüketici elektroniği ve tıbbi cihazlar gibi alanlarda bu yaklaşımın yaygınlaşması, performans iyileştirmelerinin giderek yazılım merkezli optimizasyonlara dayandığını göstermektedir (Lee vd., 2025).

Sonuç olarak, modern bilgisayar sistemlerinin gelişimi yalnızca donanım performansının artırılmasıyla açıklanamayacak kadar karmaşık bir hâl almıştır. Enerji verimliliği, güvenilirlik, ölçeklenebilirlik ve maliyet gibi faktörler, sistem tasarımında yazılımın rolünü giderek daha kritik hâle getirmiştir. Bu nedenle günümüz bilgisayar mimarilerinde donanım altyapısı ile yazılım optimizasyonlarının birlikte ele alındığı bütünleşik tasarım yaklaşımları ön plana çıkmaktadır (Lee vd., 2025). Gelecekte AI destekli sistem tasarımı, donanım soyutlama katmanları ve modüler yazılım mimarileri gibi yaklaşımların, donanım sınırlamalarını dengeleyen temel optimizasyon araçları olarak daha da önem kazanması beklenmektedir.

1.3. Veri Yoğun Dünyada Hesaplama Gereksinimleri

Günümüzde üretilen veri miktarı tarihsel olarak görülmemiş bir hızla artmaktadır. İnternet, mobil cihazlar, sensör ağları ve IoT sistemleri tarafından üretilen veri akışı; metin, görüntü, video ve sensör verileri gibi farklı türlerde büyük veri ekosistemlerini ortaya çıkarmıştır. Bu gelişmeler, geleneksel bilgi işlem mimarilerinin kapasitesini zorlamakta ve veri yoğun (data-intensive) hesaplama paradigmalarının önemini artırmaktadır. Veri yoğun hesaplama; büyük veri kümelerinin toplanması, depolanması, işlenmesi ve analiz edilmesini kapsayan çok aşamalı bir süreç olup özellikle makine öğrenmesi, grafik analizi ve veri madenciliği gibi yöntemler bu süreçte temel rol oynamaktadır. Bu nedenle modern bilgi işlem sistemlerinin yalnızca hesaplama gücü değil, aynı zamanda veri erişimi ve veri işleme kapasitesi açısından da ölçeklenebilir olması gerekmektedir (Akarvardar ve Wong, 2023).

Büyük veri çağında hesaplama gereksinimlerinin artmasının en önemli nedenlerinden biri, veri hacminin (volume), veri üretim hızının (velocity) ve veri çeşitliliğinin (variety) sürekli büyümesidir. Örneğin küresel veri evreninin 2020-2025 yılları arasında yaklaşık 3,5 kat artacağı öngörülmektedir (Evans vd., 2024). Bu hızlı büyüme, veri analizi süreçlerinin yalnızca depolama ve erişim açısından değil, aynı zamanda yüksek bant genişliği ve paralel işlem kapasitesi açısından da yeni altyapılar gerektirmesine yol açmaktadır. Veri yoğun uygulamalar genellikle milyonlarca basit işlemin eş zamanlı olarak yürütülmesini gerektirdiğinden GPU, FPGA ve özel amaçlı hızlandırıcılar gibi paralel mimarilere dayalı donanımlar bu alanda önemli bir rol oynamaktadır.

Geleneksel süper bilgisayar sistemleri uzun süre yalnızca hesaplama yoğun (compute-intensive) bilimsel problemlerin çözümüne odaklanmış olsa da günümüzde veri yoğun uygulamalar HPC altyapılarının temel kullanım alanlarından biri hâline gelmiştir. Özellikle iklim modelleme, genom analizi, mühendislik tasarımı ve AI uygulamaları gibi alanlarda büyük veri analitiği ile HPC sistemlerinin entegrasyonu kritik bir ihtiyaç hâline gelmiştir. Bu entegrasyon sayesinde çok büyük veri kümeleri paralel işlem teknikleri kullanılarak kısa sürede analiz edilebilmekte ve karmaşık problemlere yönelik yeni çözümler geliştirilebilmektedir (Pyzer-Knapp vd., 2022).

Son yıllarda veri yoğun uygulamaların yaygınlaşmasıyla birlikte süper bilgisayar mimarilerinin tasarım hedefleri de değişmeye başlamıştır. Güncel HPC sistemleri yalnızca işlem gücünü artırmayı değil, aynı zamanda veri işleme kapasitesini, depolama altyapısını ve veri aktarım hızını da optimize etmeyi amaçlamaktadır (Navaux vd., 2023). Büyük veri analitiği ve AI uygulamalarının HPC ortamlarında yaygınlaşması, veri işleme süreçlerinin sistem performansındaki en kritik darboğazlardan biri hâline gelmesine neden olmuştur. Bu nedenle modern süper bilgisayar mimarileri; yüksek bant genişliğine sahip depolama sistemleri, paralel veri işleme mekanizmaları ve veri merkezleri ile entegre çalışan veri-yoğun altyapılar geliştirmeye yönelmektedir.

1.4. Disiplinler Arası Dönüşüm (Sağlık, Enerji, Finans, Savunma)

2010 sonrasında bilgisayar bilimlerinde yaşanan teknolojik gelişmeler yalnızca bilişim alanını değil, aynı zamanda farklı sektörleri de derinden etkileyen disiplinler arası bir dönüşümü beraberinde getirmiştir. Büyük veri analitiği, AI, HPC ve bulut bilişim gibi teknolojiler; sağlık, enerji, finans ve savunma gibi kritik sektörlerde karar verme süreçlerini yeniden şekillendirmiştir. Bu dönüşüm, veri odaklı yaklaşımların farklı disiplinlerle entegrasyonunu mümkün kılarak daha karmaşık problemlerin çözülebilmesine olanak sağlamaktadır (Egger ve Yu, 2022). Böylece bilgisayar bilimleri, yalnızca teknik bir alan olmaktan çıkarak çok disiplinli araştırmaların merkezinde yer alan stratejik bir bilim dalı hâline gelmiştir.

Sağlık alanında bu dönüşüm özellikle AI destekli tanı sistemleri, tıbbi görüntü analizi ve kişiselleştirilmiş tıp uygulamaları ile belirginleşmiştir. Derin öğrenme tabanlı algoritmalar, radyoloji görüntülerinde tümör tespiti, patoloji görüntülerinde hücresel yapıların analizi ve genom verilerinin yorumlanması gibi alanlarda yüksek doğruluk oranları sağlayarak klinik karar destek sistemlerinin gelişimine katkı sunmaktadır (Thirunavukarasu ve Kotei, 2024). Bunun yanında büyük veri analitiği sayesinde hastane bilgi sistemlerinden, biyomedikal sensörlerden ve genomik çalışmalardan elde edilen veriler entegre

edilerek hastalıkların erken teşhisi ve tedavi süreçlerinin optimize edilmesi mümkün hâle gelmiştir.

Enerji sektöründe ise bilgisayar bilimlerinin sunduğu veri analitiği ve AI teknikleri, enerji üretim ve dağıtım süreçlerinin daha verimli hâle getirilmesinde önemli rol oynamaktadır. Akıllı şebekeler (smart grids), sensör ağları ve IoT tabanlı enerji izleme sistemleri sayesinde enerji tüketimi gerçek zamanlı olarak analiz edilebilmekte ve üretim planlaması daha etkin biçimde yapılabilmektedir. Özellikle yenilenebilir enerji kaynaklarının entegrasyonu, enerji talep tahmini ve şebeke optimizasyonu gibi konularda makine öğrenmesi algoritmaları önemli avantajlar sağlamaktadır (Maheshwari vd., 2022). Bu sayede enerji sistemleri daha sürdürülebilir, güvenilir ve verimli bir yapıya kavuşmaktadır.

Finans ve savunma sektörleri de bilgisayar bilimlerindeki ilerlemelerden önemli ölçüde etkilenmiştir. Finans alanında algoritmik ticaret sistemleri, dolandırıcılık tespit mekanizmaları ve risk analizi modelleri büyük veri ve makine öğrenmesi teknikleri ile geliştirilmektedir (Kumar vd., 2025). Bu sistemler milyonlarca finansal işlemi gerçek zamanlı olarak analiz ederek piyasa davranışlarını modelleyebilmekte ve yatırım kararlarını destekleyebilmektedir. Savunma alanında ise büyük veri analitiği, siber güvenlik, otonom sistemler ve karar destek platformları modern askeri operasyonların önemli bileşenleri hâline gelmiştir. Özellikle AI destekli gözetleme sistemleri, tehdit analizi ve stratejik planlama araçları savunma teknolojilerinde yeni bir paradigma oluşturmuştur (Weber vd., 2024). Bu gelişmeler, bilgisayar bilimlerinin farklı disiplinlerle bütünleşerek küresel ölçekte stratejik bir dönüşüm yarattığını göstermektedir.

Savunma sanayi alanında bilgisayar bilimleri ve özellikle AI teknolojileri, modern askeri stratejilerin önemli bileşenlerinden biri hâline gelmiştir. AI ve makine öğrenmesi tabanlı sistemler; istihbarat toplama, gözetleme, keşif (ISR), siber güvenlik ve hedef tespiti gibi kritik askeri faaliyetlerde kullanılmaktadır. Bu teknolojiler sayesinde büyük veri kümeleri hızlı biçimde analiz edilerek savaş alanındaki durumsal farkındalık artırılmakta ve komuta kontrol süreçlerinde daha hızlı ve doğru kararlar alınabilmektedir. Özellikle otonom sistemler, insansız hava araçları ve akıllı savunma platformları gibi teknolojiler askeri operasyonların etkinliğini artıran önemli araçlar olarak öne çıkmaktadır (Khan vd., 2021).

Günümüzde ABD, Çin ve Rusya gibi teknolojik açıdan gelişmiş ülkeler savunma sistemlerine AI tabanlı yetenekler entegre etmeye yönelik yoğun yatırımlar yapmaktadır. Bu çerçevede geliştirilen otonom silah sistemleri, insan-makine iş birliği ile çalışan karar destek mekanizmaları ve ağ tabanlı savunma platformları, modern savaş doktrinlerinde giderek daha fazla yer almaktadır. Ayrıca AI destekli sistemler, yüksek hızda veri işleyebilme kabiliyeti sayesinde

savaş alanında insanın reaksiyon süresinin ötesinde karar alma süreçlerine katkı sağlayabilmektedir (Khan vd., 2021; Weng, 2024). Bu durum, gelecekte askeri güç dengesinin yalnızca geleneksel silah sistemleriyle değil, aynı zamanda gelişmiş yazılım ve algoritmik yeteneklerle de belirleneceğini göstermektedir.

2. Üretken AI ve Temel Modeller

Üretken AI ve temel modeller (Foundation Models), son yıllarda AI araştırmalarında ortaya çıkan en önemli paradigmalardan birini temsil etmektedir. Üretken AI; metin, görüntü, ses ve video gibi yeni içerikler üretebilen ve büyük veri kümeleri üzerinde eğitilmiş derin öğrenme modellerine dayanan sistemleri ifade eder. Bu sistemler, özellikle transformer tabanlı mimariler sayesinde dil üretimi, görüntü sentezi, kod üretimi ve çok modlu içerik oluşturma gibi görevlerde insan benzeri çıktılar üretebilmektedir. Temel modeller ise çok büyük ölçekli veri setleri üzerinde önceden eğitilmiş ve farklı görevler için yeniden uyarlanabilen genel amaçlı AI modelleridir. Bu modeller, transfer öğrenme ve ince ayar (fine-tuning) yöntemleri sayesinde doğal dil işleme, bilgisayarlı görü ve konuşma işleme gibi birçok farklı uygulama alanında kullanılabilmektedir (Fui-Hoon Nah vd., 2023; Waqas vd., 2023). Günümüzde üretken AI sistemlerinin önemli bir bölümü bu temel modeller üzerine inşa edilmekte olup, bu yaklaşım AI geliştirme süreçlerini hızlandırmakta ve farklı disiplinlerdeki uygulamaların daha ölçeklenebilir ve erişilebilir hâle gelmesini sağlamaktadır.

2.1. Büyük Dil Modelleri (LLM) ve Çok Modlu Sistemler

Büyük Dil Modelleri (LLM'ler), doğal dili yalnızca istatistiksel örüntüler düzeyinde değil, bağlam, anlam ilişkileri ve görev niyeti düzeyinde de işleyebilen, çoğunlukla Transformer temelli ve milyarlarca parametreyle eğitilmiş temel modellerdir (Chen vd., 2024). Bu modeller; çok büyük metin koleksiyonları üzerinde ön eğitimden geçirilerek metin üretimi, çeviri, özetleme ve soru-cevaplama gibi görevlerde yüksek başarı göstermekte, ayrıca ölçek büyüdükçe bağlam içi öğrenme, yönerge izleme ve çok adımlı akıl yürütme gibi daha önce küçük modellerde belirgin olmayan "ortaya çıkan yetenekler" sergileyebilmektedir. Bu nedenle LLM'ler, yalnızca klasik doğal dil işleme araçları olarak değil; bilgiye dayalı karar destek, etkileşimli yardımcı sistemler ve genel amaçlı AI ajanlarının bilişsel çekirdeği olarak da değerlendirilmektedir. Nitekim son dönem çalışmalar, GPT, LLaMA ve PaLM gibi ailelerin LLM ekosistemini şekillendirdiğini; bu modellerin geniş ölçekli ön eğitim, ince ayar ve hizalama süreçleri sayesinde dil anlama ve üretme kapasitesini belirgin biçimde ileri taşıdığını göstermektedir (Chen vd., 2024; Minaee vd., 2024).

Çok modlu sistemler ise LLM'lerin bu dilsel yeteneklerini metin dışındaki veri türleriyle birleştirerek görüntü, ses, video ve sensör çıktıları gibi heterojen girdileri ortak bir anlamsal uzayda işleyebilen daha kapsamlı yapılar ortaya koymaktadır. Bu çerçevede büyük çok modlu modeller (LMM/MLM), yalnızca görsel betimleme ya da görsel soru-cevap gibi görevleri yerine getirmekle kalmayıp, farklı modaliteler arasındaki ilişkileri öğrenerek çapraz-modal akıl yürütme, çok modlu diyalog ve bağlama duyarlı karar verme yetenekleri geliştirmektedir. Wang ve arkadaşlarına (2024) göre, KOSMOS-1, Gemini, BLIP-2, Flamingo, MiniGPT-4 ve LLaVA gibi örnekler; metin ile görüntü arasındaki hizalamanın, dondurulmuş görsel kodlayıcılar ile LLM'lerin ara katmanlar üzerinden bütünleştirilmesinin ve çok modlu ön eğitim stratejilerinin bu dönüşümde merkezi rol oynadığını göstermektedir. Dahası, otonom sürüşte LiDAR, kamera ve dil girdilerinin birlikte yorumlanması ya da insan-robot etkileşiminde konuşma, bakış, duruş ve nesne konumlarının doğal dil üzerinden işlenmesi, çok modlu sistemlerin fiziksel dünyaya ilişkin daha zengin ve uygulanabilir bir AI anlayışı sunduğunu ortaya koymaktadır (Wang vd., 2024; Li vd., 2025). Bu nedenle çok modlu sistemler, LLM'lerin metin merkezli sınırlarını aşarak üretken AI'yı daha kapsayıcı, etkileşimli ve gerçek dünya ile uyumlu bir yöne taşımaktadır.

2.2. Edge AI ve Hafif Model Mimarileri

Edge AI, AI modellerinin veriyi uzak bulut merkezlerine taşımadan doğrudan cihaz üzerinde ya da ağın kenarındaki düğümlerde çalıştırılmasını ifade eder. Bu yaklaşımın temel gerekçesi; gecikmeyi azaltmak, gerçek zamanlı karar vermeyi mümkün kılmak, bant genişliği kullanımını düşürmek ve özellikle sağlık, endüstriyel IoT, akıllı şehirler, otonom sistemler ve siber güvenlik gibi alanlarda yerel veri işleme avantajı sağlamaktır. Ancak kenar cihazlar; işlem gücü, bellek, enerji tüketimi ve ağ erişimi bakımından bulut altyapılarına göre belirgin sınırlılıklara sahiptir. Bu nedenle Edge AI ekosisteminde başarı, yalnızca güçlü modeller geliştirmekle değil; sınırlı kaynaklarda kabul edilebilir doğruluk, düşük gecikme ve enerji verimliliği arasında dengeli bir mimari kurabilmekle ilişkilidir (Singh ve Gill, 2023). Edge ortamları; heterojen donanımlar, dağıtık yapı, sınırlı CPU/bellek kapasitesi ve zaman zaman kesintili bağlantılar nedeniyle klasik büyük AI modelleri için uygun değildir; bu yüzden model sıkıştırma, yerel çıkarım ve edge-to-cloud hibrit mimariler temel tasarım ilkeleri hâline gelmiştir.

Bu bağlamda hafif model mimarileri, Edge AI'ın uygulanabilirliğini sağlayan temel yapı taşlarıdır. MobileNet ailesi, derinlik ayrılabilir evrişimler sayesinde işlem yükünü ve parametre sayısını azaltırken; SqueezeNet, "fire module" tasarımıyla çok daha küçük model boyutlarında rekabetçi performans

sunmaktadır. TinyML yaklaşımı ise mikrodenetleyici sınıfı cihazlarda dahi çıkarımı mümkün kılarak ultra düşük güç tüketimli uygulamaların önünü açmaktadır. Bunun yanında budama (pruning), nicemleme (quantization) ve bilgi damıtma (knowledge distillation) gibi sıkıştırma teknikleri, hafif modellerin edge cihazlara dağıtılabilirliğini daha da güçlendirmektedir (Babalola vd., 2024). Güncel çalışmalarda yalnızca klasik hafif CNN'ler değil; EfficientNetV2 türevleri, MobileViTv2, EdgeViTs ve EdgeNeXt gibi yeni nesil mimariler de öne çıkmakta; hatta örnek bir çalışmada geliştirilen "Linge" modeli yalnızca 7.63 MB parametre boyutuyla edge sunucular üzerinde çalıştırılabilirken yüksek doğruluk ve AUC değerleri elde etmiştir. Bu durum, hafif mimarilerin artık yalnızca "küçük model" değil; bellek erişim maliyetini azaltan, dikkat mekanizmalarını seçici biçimde kullanan ve gerçek dünya edge senaryolarına göre yeniden tasarlanan özel mimari çözümler olarak ele alınması gerektiğini göstermektedir (Zhou vd., 2024).

2.3. Açıklanabilir AI (Explainable AI-XAI)

XAI, özellikle derin öğrenme ve diğer karmaşık makine öğrenmesi modellerinin "kara kutu" niteliğini azaltmayı; modelin hangi veri, örüntü, özellik ya da mantıksal ilişki üzerinden sonuca ulaştığını insanlar için anlaşılır hâle getirmeyi amaçlayan yöntemler bütünüdür (Bilal vd., 2025). Bu yönüyle XAI, yalnızca teknik bir şeffaflık aracı değil; aynı zamanda güven, hesap verebilirlik ve düzenleyici uyum açısından kritik bir çerçevedir. Angelov ve arkadaşlarının (2021) analitik incelemesinde vurgulandığı üzere, yorumlanabilirlik tek başına kara kutu modellerin doğurduğu tüm sorunları karşılamaya yetmez; kullanıcıların güvenini kazanmak ve kararların nedenlerini anlamlandırmak için açıklanabilirlik gereklidir. Aynı çalışmada NIST'in dört temel XAI ilkesi de öne çıkarılmaktadır: sistemin kararına dair bir gerekçe sunması, bu gerekçenin kullanıcı için anlamlı olması, açıklamanın sistemin gerçek işleyişini doğru yansıtması ve modelin bilgi sınırlarını tanıyabilmesi. Bu çerçeve, XAI'ın yalnızca model içi tekniklerin toplamı olmadığını; yerel ve küresel açıklamalar, post-hoc yaklaşımlar ve açıklanabilir-tasarım ilkeleri arasında kurulan bütüncül bir güven mimarisi olduğunu göstermektedir.

Güncel literatür ise XAI'ın etkili olabilmesi için açıklamaların yalnızca teknik olarak üretilmesinin yeterli olmadığını, insan-merkezli ve bağlama duyarlı olması gerektiğini göstermektedir (Budhkar vd., 2025; Nandan vd., 2025; Nikiforidis vd., 2025). Sağlık alanına odaklanan çalışmada, XAI'ın veri, muhakeme ve karar süreçlerine ilişkin içgörü sağlayarak insan anlayışını ve güveni artırdığı; hata ve önyargıların görünür kılınmasına katkı sunduğu belirtilmektedir. Bununla birlikte mevcut yaklaşımların çoğu hâlen algoritma-merkezlidir; oysa özellikle sağlık gibi yüksek riskli alanlarda açıklamanın hangi paydaş için,

hangi karar anında ve hangi etik gereksinimlerle üretildiği belirleyicidir. Bu nedenle insan-merkezli açıklanabilir AI yaklaşımı, teknik açıklamayı sosyo-teknik boyut, insan değerleri ve kullanıcı ihtiyaçlarıyla birleştirmektedir. LLM tabanlı güncel XAI yazını da bu yönelimi desteklemekte; post-hoc açıklamalar, içkin yorumlanabilirlik ve insan-merkezli anlatısal açıklamalar arasında bir ayrım yaparak, iyi bir açıklamanın yalnızca "doğru" değil, aynı zamanda anlaşılır, bağlama uygun ve kullanıcı tarafından denetlenebilir olması gerektiğini vurgulamaktadır (van Leersum vd., 2025). Dolayısıyla XAI, çağdaş AI sistemlerinde performans ile şeffaflık arasındaki gerilimi yönetmeye çalışan; güvenilir, adil ve kullanıcıyla iş birliği kurabilen sistemlerin inşasında temel rol oynayan bir yaklaşım olarak değerlendirilmelidir.

2.4. Model Güvenilirliği ve Halüsinasyon Problemi

Model güvenilirliği, AI sistemlerinin farklı bağlamlarda tutarlı, doğrulanabilir ve görevin gerektirdiği doğruluk düzeyinde çıktılar üretebilme kapasitesiyle ilişkilidir. Büyük dil modelleri bağlamında bu güvenilirlik sorunu, doğrudan model mimarisinin sınırlarıyla bağlantılıdır; çünkü LLM'ler gerçeği denetleyen bir mekanizma üzerinden değil, önceki belirteçlere dayanarak sonraki belirteci olasılıksal biçimde üreten otoregresif yapılar olarak çalışır (Ahn, 2025). Bu nedenle özellikle uzmanlık gerektiren alanlarda uzun menzilli tutarlılık, teknik doğruluk ve bağlama sadakat her zaman garanti edilememektedir. Halüsinasyon, yalnızca geçici bir mühendislik kusuru değil; kimi çalışmalara göre LLM mimarisinin yapısal bir sonucu olarak ortaya çıkmakta ve olgusal yanlışlık, yanlış yorumlama, kısmi yanlışlık ve uydurma gibi farklı biçimlerde görülebilmektedir. Ayrıca güncel değerlendirmeler, soru-cevap görevlerinde modellerin hâlen anlamlı düzeyde halüsinasyon üretme eğiliminde olduğunu göstermekte; örneğin bazı benchmark'larda bu oranların yaklaşık %17 ile %55 arasında değişebildiği raporlanmaktadır (Ahn, 2025; Yang vd., 2025). Literatürde halüsinasyon; girdiyle çelişen, bağlamla çelişen ve yerleşik dünya bilgisiyle çelişen çıktılar olarak sınıflandırılmakta, ayrıca veri toplama, eğitim ve çıkarım aşamalarındaki kalite sorunları, önyargılar, uzun bağlam işleme sınırlılıkları ve örnekleme rastlantısallığı da bu problemin başlıca nedenleri arasında gösterilmektedir (Yang vd., 2025b).

Halüsinasyon probleminin güvenilirlik üzerindeki etkisi, özellikle tıp, hukuk, bilimsel araştırma ve savunma gibi yüksek riskli alanlarda daha belirgin hâle gelmektedir; çünkü akıcı ve ikna edici görünen ama yanlış olan içerikler, uzman olmayan kullanıcıları kolayca yanıltabilmektedir. Bu nedenle güncel çalışmalar, model güvenilirliğini artırmak için çok katmanlı doğrulama ve denetim stratejileri önermektedir. İleri düzey istem mühendisliği teknikleri arasında Chain-of-Thought, Self-Consistency, Decomposition ve Chain-of-

Verification gibi yaklaşımlar, yanıtların daha sistematik üretilmesine katkı sağlarken; Retrieval-Augmented Generation (RAG) sistemleri modeli güvenilir veri tabanlarıyla destekleyerek çıktıları olgusal zemine oturtmayı amaçlamaktadır (Ahn, 2025; Yang vd., 2025). Bununla birlikte Yang ve arkadaşlarına (2025) göre, RAG'ın da hataları tamamen ortadan kaldırmadığı açıkça görülmektedir. Daha yeni yaklaşımlar arasında ise çok-etmenli tartışma ve oylama mekanizmaları, tekrar sorgulama, hata günlüğü ve öz-yansıma süreçleri ile metamorfik ilişkiler üzerinden çalışan sıfır-kaynaklı halüsinasyon tespit yöntemleri öne çıkmaktadır. Bu çalışmaların ortak sonucu, model güvenilirliğinin tek başına model ölçeğiyle çözülemeyeceğini; güvenilir kullanım için görevle uyumlu araç seçimi, sistematik doğrulama, çapraz denetim ve özellikle insan gözetiminin vazgeçilmez olduğunu göstermesidir (Yang vd., 2025). Dolayısıyla halüsinasyon problemi, yalnızca teknik bir üretim hatası değil; güven, hesap verebilirlik ve gerçek dünya uygulamalarında AI sistemlerinin sınırlarını belirleyen temel bir güvenilirlik meselesidir.

2.5. Akademik ve endüstriyel kullanım alanları

Akademik kullanım alanlarında XAI, özellikle kararın yalnızca doğru olmasının değil, neden doğru olduğunun da gösterilmesi gereken yüksek etkili araştırma alanlarında öne çıkmaktadır. Kalasampath ve arkadaşlarına (2025) göre, son dönem XAI uygulamalarının en yoğun biçimde sağlık alanında toplandığını; kanser tanısı, COVID-19 yönetimi, tıbbi görüntüleme, nörobilim ve genel klinik karar destek süreçlerinde açıklanabilirliğin güven, hesap verebilirlik ve uzman doğrulaması açısından kritik görüldüğünü göstermektedir. Bunun yanında çevre ve tarım, finans, hukuk, eğitim, sosyal bakım, ulaşım ve siber güvenlik gibi alanlarda da akademik çalışmalar hızla artmaktadır. Eğitimde öğrenci performansının öngörülmesi ve öğrenme analitikleri, sosyal medyada yanlış bilgi ve nefret söylemi analizi, hukukta karar destek, siber güvenlikte saldırı ve anomali tespiti gibi örnekler, XAI'ın yalnızca teknik bir yorumlama aracı değil; alan uzmanlarının model davranışını değerlendirmesine imkân veren disiplinler arası bir araştırma altyapısı hâline geldiğini ortaya koymaktadır. Ayrıca bu literatürde SHAP ve LIME gibi yerel açıklama yöntemlerinin baskın olduğu, görselleştirme ve özellik önem derecelendirmesinin en yaygın açıklama biçimleri arasında yer aldığı görülmektedir (Kalasampath vd., 2025; Saarela vd., 2024).

3. Güvenli ve Etik AI Sistemleri

3.1. Düşmanca (Adversarial) Saldırılar

Adversarial saldırılar, makine öğrenmesi ve özellikle derin öğrenme modellerinin giriş verisine insan tarafından çoğu zaman fark edilemeyecek kadar küçük fakat stratejik bozulmalar eklenerek hatalı karar vermeye zorlanmasıdır. Qiu ve arkadaşlarına (2019) göre, bu saldırıların eğitim aşamasında veri zehirleme, etiket manipülasyonu ve özellik bozma biçiminde; test aşamasında ise beyaz kutu ve siyah kutu saldırıları olarak ortaya çıktığı belirtilmektedir. Ayrıca bu saldırıların yalnızca görüntü sınıflandırma ile sınırlı kalmadığı; doğal dil işleme, siber güvenlik, bulut servisleri ve fiziksel dünya uygulamalarında da ciddi tehdit oluşturduğu vurgulanmaktadır. Bu durum, AI sistemlerinde yüksek doğruluk oranlarının tek başına yeterli olmadığını; model dayanıklılığı, savunma mekanizmaları ve güvenli dağıtım stratejilerinin de en az performans kadar önemli olduğunu göstermektedir (Qiu vd., 2019).

3.2. Model zehirleme (Data/Model Poisoning)

Data/model poisoning, AI sistemlerinin eğitim verisini, doğrulama kümelerini ya da doğrudan modelin mantığını kasıtlı biçimde bozarak performansı düşürmeyi veya belirli yanlış kararlar üretmesini sağlamayı amaçlayan saldırıları ifade eder. Kaynağa göre veri zehirleme; eğitim ya da doğrulama verisine yeni örnekler ekleme, mevcut içerik, etiket veya özellikleri değiştirme ya da veriyi silme yoluyla gerçekleştirilebilir ve hedefli ya da hedefsiz biçimde uygulanabilir. Model poisoning ya da mantık bozma saldırıları ise modelin algoritmasını, kodunu, gradyanlarını, kurallarını veya karar prosedürlerini değiştirerek doğruluğu azaltabilir ya da sistemi kötü niyetli çıktılar üretmeye yöneltebilir. Bu nedenle poisoning saldırıları, yalnızca veri kalitesine yönelik bir tehdit değil; veri işleme, model geliştirme ve dağıtım aşamalarının tamamını etkileyebilen, AI tabanlı yazılımların güvenilirliğini ve bütünlüğünü doğrudan zedeleyen kritik bir güvenlik sorunudur (Kumar vd., 2024).

3.3. AI Güvenliği ve Siber Güvenlik Kesişimi

AI güvenliği ile siber güvenliğin kesişimi, AI'nin yalnızca korunması gereken bir hedef değil, aynı zamanda siber savunmayı dönüştüren etkin bir araç hâline gelmesiyle ortaya çıkmaktadır. Sontan ve Samuel'e (2024) göre AI; tehdit tespiti, zafiyet analizi ve olay müdahalesi gibi alanlarda büyük veri hacimlerini yüksek hızda işleyerek anomali ve saldırı örüntülerini geleneksel yöntemlere göre daha hızlı saptayabilmekte, böylece kurumların savunma kapasitesini artırmaktadır. Bununla birlikte, bu bütünleşmenin yeni riskler de ürettiği

vurgulanmaktadır: AI sistemlerinin şeffaflık eksikliği, önyargı, veri gizliliği, etik sorunlar ve saldırganlar tarafından kötüye kullanılabilme ihtimali, AI destekli siber güvenlik çözümlerinin dikkatli biçimde tasarlanmasını zorunlu kılmaktadır. Bu nedenle AI güvenliği ile siber güvenliğin kesişimi, bir yandan akıllı otomasyon ve öngörüsel savunma fırsatları sunarken, diğer yandan güvenilirlik, açıklanabilirlik, mahremiyet ve hesap verebilirlik ilkelerini merkeze alan çok katmanlı bir güvenlik anlayışını gerekli kılan stratejik bir alan olarak değerlendirilmektedir (Sontan ve Samuel, 2024).

3.4. Regülasyonlar (AI Act, Veri Koruma Politikaları)

Regülasyonlar bağlamında AI ile veri koruma politikalarının kesişimi, teknolojik yenilik ile bireysel hakların korunması arasında denge kurma çabasına dayanmaktadır. Yanamala ve Suryadevara'ya (2023) göre, özellikle Genel Veri Koruma Yönetmeliği'nin (General Data Protection Regulation - GDPR), kişisel verilerin işlenmesi, şeffaflık, veri minimizasyonu ve hesap verebilirlik ilkeleri üzerinden küresel veri yönetişimini derinden etkilediği; buna benzer düzenlemelerin de farklı ülkelerde yaygınlaştığı vurgulanmaktadır. Aynı metin, Avrupa Birliği'nin önerdiği AI Act yaklaşımının ise özellikle yüksek riskli AI uygulamalarına yönelik yükümlülükler getirerek temel haklar, güvenlik ve etik uygunluk ekseninde yeni bir düzenleyici çerçeve oluşturduğunu belirtmektedir. Bu çerçevede regülasyonlar, yalnızca veri ihlallerini önlemeyi değil; aynı zamanda AI sistemlerinin şeffaf, adil, izlenebilir ve insan haklarıyla uyumlu biçimde geliştirilmesini sağlamayı amaçlamaktadır. Dolayısıyla AI Act ve veri koruma politikaları birlikte değerlendirildiğinde, güncel AI yönetişiminin temelini oluşturan iki tamamlayıcı sütun olarak öne çıkmaktadır (Yanamala ve Suryadevara, 2023).

3.5. Etik Çerçeveler ve Algoritmik Adalet

Etik çerçeveler, AI sistemlerinin hangi ilkelere göre geliştirilmesi ve değerlendirilmesi gerektiğini tanımlayan normatif bir yapı sunar. Prem'e göre (2023), güncel AI etik çerçevelerinin çoğunda insan özerkliği ve gözetimi, teknik sağlamlık ve güvenlik, mahremiyet ve veri yönetişimi, şeffaflık, çeşitlilik-ayrımcılık yapmama-adalet, toplumsal fayda ve hesap verebilirlik gibi ortak ilkelerin öne çıktığını göstermektedir. Ancak aynı çalışma, bu çerçevelerin çoğunlukla ilke düzeyinde kaldığını; yani "etik AI" hedeflerini tanımlasa da bu hedeflerin somut sistem tasarımına ve uygulanabilir teknik gereksinimlere nasıl dönüştürüleceğinin çoğu zaman belirsiz olduğunu vurgulamaktadır. Bu nedenle etik çerçeveler, AI yönetişimi için gerekli bir başlangıç zemini sağlasa da gerçek etki yaratabilmeleri için ilke listelerinin pratik tasarım kararları,

denetim araçları ve bağlama duyarlı uygulama mekanizmalarıyla desteklenmesi gerekmektedir.

Algoritmik adalet ise meseleyi yalnızca teknik doğruluk ya da önyargı azaltma problemi olarak değil, algoritmik kararların toplumda kimleri nasıl etkilediğini sorgulayan daha geniş bir adalet perspektifiyle ele alır. "Theorising Algorithmic Justice" çalışmasına göre (Marjanovic vd., 2022) algoritmik adalet; "adaletin konusu nedir, kim adaletin öznesidir, adaletsizlikler nasıl üretilir ve bu ihtilaflar nasıl giderilir?" soruları etrafında düşünülmelidir. Marjanovic ve arkadaşları (2025), algoritmik adaletsizliklerin yalnızca yanlış sınıflandırmadan ibaret olmadığını; maldistribution, misrecognition ve misrepresentation gibi ekonomik, sosyo-kültürel ve siyasal boyutlar taşıdığını, ayrıca datafication, black-boxing ve algoritmik inscrutability gibi süreçlerle görünmez biçimde üretilebildiğini göstermektedir. Bu bakımdan algoritmik adalet, etik ilkelerin toplumsal sonuçlarını görünür kılar; özellikle kırılgan gruplar üzerindeki etkileri, yanlış pozitiflerin yarattığı zararları ve insan gözetimi olmadan işleyen karar sistemlerinin doğurabileceği yapısal eşitsizlikleri tartışmaya açar.

4. Yüksek Performanslı ve Dağıtık Hesaplama Sistemleri

4.1. GPU/TPU Hızlandırmalı Sistemler

GPU/TPU hızlandırmalı sistemler, derin öğrenme ve büyük ölçekli veri işleme süreçlerinde ortaya çıkan yüksek paralel hesaplama gereksinimini karşılamak üzere geliştirilen temel hızlandırma altyapıları arasında yer almaktadır. Geleneksel CPU mimarileri daha genel amaçlı ve sıralı işlem odaklı çalışırken, GPU'lar binlerce çekirdekle eşzamanlı işlem yürütebilme yetenekleri sayesinde özellikle matris çarpımları, tensör işlemleri ve derin sinir ağı eğitimi gibi yoğun hesaplama gerektiren görevlerde belirgin performans üstünlüğü sağlamaktadır. TPU'lar ise tensör tabanlı işlemler için özel olarak tasarlanmış uygulamaya özgü tümleşik devreler olarak, özellikle büyük ölçekli yapay sinir ağı eğitim ve çıkarım süreçlerinde süreyi azaltma, ölçeklenebilirliği artırma ve işlem verimliliğini yükseltme açısından öne çıkmaktadır (Emmanuel vd., 2025). CPU, GPU ve TPU platformlarını karşılaştıran çalışmalar, özellikle dağıtık derin öğrenme senaryolarında GPU ve TPU'nun klasik işlemcilere göre çok daha kısa eğitim süreleri sunduğunu; bazı iş yüklerinde TPU'nun daha dengeli ve yüksek verimli bir performans sergileyebildiğini göstermektedir. Bu nedenle GPU/TPU tabanlı hızlandırma sistemleri, yalnızca AI uygulamalarının performansını artıran teknik bileşenler değil; aynı zamanda büyük modellerin eğitilebilirliğini, deneysel çalışmaların uygulanabilirliğini ve modern yüksek performanslı hesaplama altyapılarının sürdürülebilirliğini belirleyen stratejik unsurlar olarak değerlendirilmektedir (Kimm vd., 2021).

4.2. Bulut Bilişim ve Hibrit Mimariler

Bulut bilişim, işlemci, depolama ve ağ kaynaklarının internet üzerinden isteğe bağlı ve ölçeklenebilir biçimde sunulmasını mümkün kılan bir hizmet modeli olarak, özellikle esnek kaynak kullanımı, hızlı ölçeklenme ve kullanım kadar ödeme yaklaşımı sayesinde modern bilgi işlem altyapılarının temelini oluşturmuştur (Aktas, 2018). Ancak tekil bulut yapıları zamanla sağlayıcı bağımlılığı, entegrasyon güçlüğü, güvenlik kaygıları ve farklı iş yüklerinin aynı ortamda verimli biçimde yönetilememesi gibi sınırlılıkları görünür kılmıştır. Bu nedenle hibrit mimariler, özel ve genel bulut kaynaklarını bir araya getirerek hem hassas verilerin daha kontrollü ortamlarda tutulmasına hem de değişken iş yüklerinin dış kaynaklarla elastik biçimde desteklenmesine olanak tanıyan daha dengeli bir çözüm olarak öne çıkmıştır. Bu yaklaşım, özellikle ölçeklenebilirlik, maliyet optimizasyonu, iş sürekliliği ve mevzuata uyum gereksinimlerinin birlikte ele alınması gereken kurumsal yapılarda önem kazanmaktadır (Cherukuri, 2019).

Hibrit mimarilerin asıl gücü, yalnızca farklı altyapıları birleştirmesinde değil, bu bileşenleri yönetilebilir ve birlikte çalışabilir bir yapıya dönüştürmesinde yatmaktadır. Literatürde bu amaçla önerilen mimariler; gerçek zamanlı izleme, olay akışı işleme, otomatik ölçekleme, önleyici bakım ve iş yükü yönlendirme gibi işlevleri ortak bir katmanda toplayarak bulut platformundan bağımsız yönetim anlayışını desteklemektedir (Mateescu vd., 2011). HPC ile bulutun kesiştiği yapılarda ise "elastic cluster" benzeri modeller, yerel yüksek performanslı kaynakları gerektiğinde bulut kaynaklarıyla genişleterek öngörülebilir yürütme, kapasite artışı ve cloud bursting olanağı sağlamaktadır. Böylece hibrit mimariler, bir yandan kurumsal BT'de esneklik ve güvenliği birlikte sunarken, diğer yandan bilimsel hesaplama ve büyük ölçekli iş yüklerinde performans ile kaynak verimliliği arasında işlevsel bir köprü kurmaktadır (Mateescu vd., 2011; Aktas, 2018).

4.3. HPC Kümeleri ve Slurm Tabanlı Kaynak Yönetimi

HPC kümeleri, çok sayıda işlem düğümünün yüksek hızlı ağlar üzerinden bir araya getirilerek büyük ölçekli ve yoğun hesaplama gerektiren iş yüklerinin paralel biçimde yürütülmesini sağlayan altyapılardır. Bu yapılar; CPU, GPU, bellek ve depolama gibi heterojen kaynakların birlikte kullanılmasına imkân vererek genomik analiz, görüntü işleme, bilimsel simülasyon ve büyük model çıkarımı gibi alanlarda hesaplama süresini anlamlı biçimde azaltmaktadır. Bu tür ortamlarda temel sorun, mevcut kaynakların çok kullanıcılı ve çok işli senaryolarda verimli, adil ve ölçeklenebilir biçimde tahsis edilmesidir. Slurm tabanlı kaynak yönetimi bu noktada, düğüm, CPU, GPU, RAM ve iş kuyruğu

yönetimini merkezi bir planlama mantığıyla düzenleyerek HPC altyapısının işletim omurgasını oluşturmaktadır. Slurm'un Linux tabanlı kurulum, kimlik doğrulama için Munge bağımlılığı, slurm.conf üzerinden denetleyici düğüm ve adres tanımları, servislerin yeniden başlatılması ve sbatch ile iş gönderimi gibi bileşenleri, dağıtık hesaplama kaynaklarının ortak bir hizmet yapısı altında yönetilmesini mümkün kılmaktadır (Curia-Alcantara vd., 2024).

Güncel çalışmalarda (Doosthosseini vd., 2024; Decker vd., 2025) Slurm yalnızca klasik toplu iş planlayıcısı olarak değil, aynı zamanda büyük dil modelleri gibi modern AI iş yüklerini çok düğümlü HPC ortamlarında ölçeklenebilir biçimde yürütmeyi sağlayan bir orkestrasyon katmanı olarak da ele alınmaktadır. Özellikle heterojen kümelerde CPU, GPU ve bellek tahsisinin dinamik olarak yapılması, konteyner tabanlı mikro servislerin entegrasyonu, REST tabanlı çıkarım uç noktalarının oluşturulması ve yük dengeleme mekanizmalarının devreye alınması, Slurm'un çağdaş HPC mimarilerindeki rolünü genişletmiştir. Nitekim ölçeklenebilir LLM çıkarım mimarilerini inceleyen çalışma, küçük modellerin daha yüksek eşzamanlı istekleri düşük gecikmeyle karşılayabildiğini, büyük modellerde ise doygunluk noktasına çok daha erken ulaşıldığını göstermekte; bu durum da kaynak planlama, kuyruk yönetimi, yeniden sıraya alma ve hata toleransı gibi Slurm işlevlerinin performans kadar kritik olduğunu ortaya koymaktadır (Luiz vd., 2025). Böylece Slurm tabanlı kaynak yönetimi, HPC kümelerinde yalnızca işlem dağıtımı yapan bir yazılım bileşeni değil; performans, ölçeklenebilirlik, güvenilirlik ve hizmet sürekliliğini birlikte yöneten stratejik bir yönetim katmanı hâline gelmektedir.

4.4. Edge-Cloud İş Bölümü

Edge-Cloud iş bölümü, AI ve veri yoğun uygulamalarda görevlerin gecikme, bant genişliği, enerji tüketimi ve hesaplama gereksinimlerine göre kenar ve bulut katmanları arasında paylaştırılmasını ifade eder. Bu yapıda bulut tarafı, büyük veri birikimi, yüksek işlem gücü ve model eğitimi gibi hesaplama yoğun süreçler için uygun bir ortam sunarken; edge katmanı, düşük gecikme gerektiren gerçek zamanlı çıkarım, yerel karar verme ve bağlama duyarlı hizmetler açısından kritik rol oynamaktadır (Yao vd., 2022). Literatür, tek başına bulutun büyük veri aktarımı ve gecikme sorunları nedeniyle, tek başına edge'nin ise sınırlı işlem gücü, bellek ve enerji kaynakları nedeniyle tüm gereksinimleri karşılamakta yetersiz kaldığını göstermektedir (Ma vd., 2025). Bu nedenle edge-cloud iş bölümü, özellikle gecikmeye duyarlı görevlerin edge'de, yüksek hesaplama maliyetli eğitim, koordinasyon ve küresel optimizasyon işlemlerinin ise bulutta yürütülmesine dayanan tamamlayıcı bir mimari anlayış olarak öne çıkmaktadır. Ayrıca çok atlamalı görev aktarımı, heterojen kaynak yapısı, ağ topolojisi

ve görev türlerindeki çeşitlilik, bu iş bölümünün sabit değil; dinamik görev zamanlama, kaynak ticareti, hesaplama offloading'i ve dijital ikiz destekli izleme mekanizmalarıyla sürekli optimize edilen bir süreç olduğunu göstermektedir. Böylece edge-cloud iş bölümü, yalnızca teknik bir dağıtım yaklaşımı değil; endüstriyel IoT, akıllı hizmetler ve modern AI sistemlerinde performans, ölçeklenebilirlik ve kaynak verimliliği arasında denge kuran temel bir mimari ilke hâline gelmektedir (Li vd., 2024).

4.5. Enerji Verimli Hesaplama (Green Computing)

Green Computing, yüksek performanslı bilgi işlem ve AI ekosistemlerinde artan işlem gücü ihtiyacının çevresel etkilerini azaltmayı amaçlayan bir yaklaşım olarak öne çıkmaktadır. Bu çerçevede Green AI, yalnızca model doğruluğunu değil, eğitim ve çıkarım süreçlerinin enerji tüketimi, karbon ayak izi ve kaynak verimliliği boyutlarını da değerlendirmeyi gerekli kılar. Sistematik incelemeler, özellikle 2020 sonrasında bu alandaki çalışmaların hızla arttığını; enerji tüketiminin izlenmesi, hiper parametre ayarlaması, model kıyaslama, dağıtım stratejileri ve doğruluk-enerji dengesi gibi başlıkların öne çıktığını göstermektedir (Verdecchia vd., 2023). Bulut bilişim tarafında ise veri merkezlerinin sürekli çalışan sunucular, soğutma sistemleri ve depolama altyapıları nedeniyle ciddi elektrik tüketimi ve karbon salımı ürettiği; bu nedenle akıllı iş yükü planlaması, dinamik kaynak tahsisi, otomatik ölçekleme, veri sıkıştırma, aynı verinin tekrar eden kopyalarını tespit edip tekilleştirme işlemi (deduplikasyon) ve AI destekli soğutma mekanizmalarının enerji verimliliği açısından kritik olduğu anlaşılmaktadır. Dolayısıyla enerji verimli hesaplama, yalnızca donanım tasarrufu sağlayan teknik bir optimizasyon alanı değil; algoritma tasarımı, yazılım mimarisi, veri merkezi yönetimi ve sürdürülebilir bulut altyapılarının birlikte ele alındığı bütüncül bir dönüşüm alanı olarak değerlendirilmelidir (Oloruntoba vd., 2024).

Kaynaklar

Bibri, S. E. (2019). On the sustainability of smart and smarter cities in the era of big data: an interdisciplinary and transdisciplinary literature review. *Journal of Big Data*, *6*(1), 25.

Chinnaiyan, B., Balasubaramanian, S., Jeyabalu, M., & Warrier, G. S. (2025). AI Applications–Computer Vision and Natural Language Processing. *Model optimization methods for efficient and edge AI: Federated learning architectures, frameworks and applications*, 25-41.

Zhou, P., Wang, L., Liu, Z., Hao, Y., Hui, P., Tarkoma, S., & Kangasharju, J. (2024). A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*.

Deelman, E., Dongarra, J., Hendrickson, B., Randles, A., Reed, D., Seidel, E., & Yelick, K. (2025). High-performance computing at a crossroads. *Science*, *387*(6736), 829-831.

Altukhi, Z. M., & Pradhan, S. (2025). Systematic literature review: Explainable AI definitions and challenges in education. *arXiv preprint arXiv:2504.02910*.

Keller, J. M. (2025). Architectural and Software-Based Fault Tolerance in Multicore and Lockstep Processing Systems: A Comprehensive Reliability-Centric Analysis. *Academic Reseach Library for International Journal of Computer Science & Information System*, *10*(11), 103-108.

De Souza, A. D. C., & de Freitas, H. C. (2025). A Performance Analysis of System Rollback Techniques. In *Simpósio em Sistemas Computacionais de Alto Desempenho (SSCAD)* (pp. 9-16). SBC.

Barwasser, A., Lentes, J., Riedel, O., Zimmermann, N., Dangelmaier, M., & Zhang, J. (2023). Method for the development of Software-Defined Manufacturing equipment. *international journal of production research*, *61*(19), 6467-6484.

Lee, J., Kim, J., Park, S. J., Song, B., & Moon, S. K. (2025). Software-Defined Product Architecture: Status, Challenges, and Future Perspectives. In *2025 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1088-1093). IEEE.

Akarvardar, K., & Wong, H. S. P. (2023). Technology prospects for data-intensive computing. *Proceedings of the IEEE*, *111*(1), 92-112.

Evans, M. R., Oliver, D., Zhou, X., & Shekhar, S. (2024). Spatial big data: Case studies on volume, velocity, and variety. In *Big Data* (pp. 115-138). CRC Press.

Pyzer-Knapp, E. O., Pitera, J. W., Staar, P. W., Takeda, S., Laino, T., Sanders, D. P., ... & Curioni, A. (2022). Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, *8*(1), 84.

Navaux, P. O. A., Lorenzon, A. F., & da Silva Serpa, M. (2023). Challenges in high-performance computing. *Journal of the Brazilian Computer Society*, *29*(1), 51-62.

Egger, R., & Yu, J. (2022). Data science and interdisciplinarity. *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, 35-49.

Thirunavukarasu, R., & Kotei, E. (2024). A comprehensive review on transformer network for natural and medical image analysis. *Computer Science Review*, *53*, 100648.

Maheshwari, S., Shetty, S., Ratnakar, R., & Sanyal, S. (2022). Role of computational science in materials and systems design for sustainable energy applications: An industry perspective. *Journal of the Indian Institute of Science*, *102*(1), 11-37.

Kumar, S., Sharma, D., Rao, S., Lim, W. M., & Mangla, S. K. (2025). Past, present, and future of sustainable finance: insights from big data analytics through machine learning of scholarly research. *Annals of Operations Research*, *345*(2), 1061-1104.

Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of Explainable Artificial Intelligence in Finance-a systematic review of Finance, Information Systems, and Computer Science literature: P. Weber et al. *Management Review Quarterly*, *74*(2), 867-907.

Khan, A., Imam, I., & Azam, A. (2021). Role of Artificial Intelligence in Defence Strategy. *Strategic Studies*, *41*(1), 19-40.

Weng, Y. (2024). Big data and machine learning in defence. *International Journal of Computer Science and Information Technology*, *16*(2), 25-35.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of information technology case and application research*, *25*(3), 277-304.

Waqas, A., Bui, M. M., Glassy, E. F., El Naqa, I., Borkowski, P., Borkowski, A. A., & Rasool, G. (2023). Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Laboratory investigation*, *103*(11), 100255.

Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... & Feng, H. (2024). Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Computers, Materials & Continua*, *80*(2).

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Wang, C., Hasler, S., Tanneberg, D., Ocker, F., Joublin, F., Ceravola, A., ... & Gienger, M. (2024, May). Lami: Large language models for multi-modal

human-robot interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-10).

Li, J., Li, J., Yang, G., Yang, L., Chi, H., & Yang, L. (2025). Applications of large language models and multimodal large models in autonomous driving: A comprehensive review. *Drones*, *9*(4), 238.

Singh, R., & Gill, S. S. (2023). Edge AI: a survey. *Internet of Things and Cyber-Physical Systems*, *3*, 71-92.

Babalola, O., Raji, O. M. O., Akande, J. O., Abdulkareem, A. O., Anyah, V., Samson, A., & Folorunso, S. (2024). AI-powered cybersecurity in edge computing: Lightweight neural models for anomaly detection. *International Journal of Multidisciplinary Research and Growth Evaluation*, *5*(2), 1130-1138.

Zhou, F., Hu, S., Du, X., Wan, X., & Wu, J. (2024). A lightweight neural network model for disease risk prediction in edge intelligent computing architecture. *Future Internet*, *16*(3), 75.

Bilal, A., Ebert, D., & Lin, B. (2025). Llms for explainable ai: A comprehensive survey. *arXiv preprint arXiv:2504.00125*.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.

Budhkar, A., Song, Q., Su, J., & Zhang, X. (2025). Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics. *Computational and Structural Biotechnology Journal*, *27*, 346-359.

Nandan, M., Mitra, S., & De, D. (2025). GraphXAI: a survey of graph neural networks (GNNs) for explainable AI (XAI). *Neural Computing and Applications*, *37*(17), 10949-11000.

Nikiforidis, K., Kyrtsoglou, A., Vafeiadis, T., Kotsiopoulos, T., Nizamis, A., Ioannidis, D., ... & Sarigiannidis, P. (2025). Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of industry 4.0/5.0. *ICT Express*, *11*(1), 135-148.

van Leersum, C. M., & Maathuis, C. (2025). Human centred explainable AI decision-making in healthcare. *Journal of Responsible Technology*, *21*, 100108.

Ahn, S. (2025). A guide to evade hallucinations and maintain reliability when using large language models for medical research: a narrative review. *Annals of Pediatric Endocrinology & Metabolism*, *30*(3), 115-118.

Yang, B., Mamun, M. A. A., Zhang, J. M., & Uddin, G. (2025). Hallucination detection in large language models with metamorphic relations. *URL https://arxiv. org/abs/2502.15844*.

Yang, Y., Ma, Y., Feng, H., Cheng, Y., & Han, Z. (2025). Minimizing hallucinations and communication costs: Adversarial debate and voting mechanisms in llm-based multi-agents. *Applied Sciences*, *15*(7), 3676.

Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Maruthamuthu, A. (2025). A literature review on applications of explainable artificial intelligence (XAI). *IEEE access*, *13*, 41111-41140.

Saarela, M., & Podgorelec, V. (2024). Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, *14*(19), 8884.

Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, *9*(5), 909.

Kumar, V., Mayo, J., & Bahiss, K. (2024). Admin: Attacks on dataset, model and input. a threat model for ai based software. *arXiv preprint arXiv:2401.07960*.

Sontan, A. D., & Samuel, S. V. (2024). The intersection of Artificial Intelligence and cybersecurity: Challenges and opportunities. *World Journal of Advanced Research and Reviews*, *21*(2), 1720-1736.

Yanamala, A. K. Y., & Suryadevara, S. (2023). Advances in data protection and artificial intelligence: Trends and challenges. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(01), 294-319.

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, *3*(3), 699-716.

Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2022). Theorising algorithmic justice. *European Journal of Information Systems*, *31*(3), 269-287.

Emmanuel, F. C., Henry, O. N., & Chibuzo, O. B. (2025). A survey comparing specialized hardware and evolution in cpu, gpu and tpu for neural network. *IRE Journals*, *8*.

Kimm, H., Paik, I., & Kimm, H. (2021). Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning. In *2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)* (pp. 312-319). IEEE.

Aktas, M. S. (2018). Hybrid cloud computing monitoring software architecture. *Concurrency and Computation: Practice and Experience*, *30*(21), e4694.

Cherukuri, B. R. (2019). Future of cloud computing: Innovations in multi-cloud and hybrid architectures. *World J. Adv. Res. Rev*, *1*(1), 068-081.

Mateescu, G., Gentzsch, W., & Ribbens, C. J. (2011). Hybrid computing-where HPC meets grid and cloud computing. *Future Generation Computer Systems*, *27*(5), 440-453.

Curia-Alcantara, N. E., Inga-Coveñas, C., & Auccahuasi, W. (2024). Methodology for the Implementation of Slurm-based HPC Services. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 771-775). IEEE.

Doosthosseini, A., Decker, J., Nolte, H., & Kunkel, J. M. (2024). Chat ai: A seamless slurm-native solution for hpc-based services. *arXiv preprint arXiv:2407.00110*.

Decker, J., Metje, S., & Kunkel, J. (2025). Running Kubernetes workloads on rootless HPC systems using Slurm. *CLOUD COMPUTING*, *111*, 2025.

Luiz, A. D. L., Kurlekar, S. V., & Georges, M. (2025). Scalable Engine and the Performance of Different LLM Models in a SLURM based HPC architecture. *arXiv preprint arXiv:2508.17814*.

Yao, J., Zhang, S., Yao, Y., Wang, F., Ma, J., Zhang, J., ... & Yang, H. (2022). Edge-cloud polarization and collaboration: A comprehensive survey for AI. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7), 6866-6886.

Ma, Q., Qin, Y., Zhu, C., Gao, L., & Chen, X. (2025). Joint resource trading and task scheduling in edge-cloud computing networks. *IEEE Transactions on Networking*, *33*(3), 994-1008.

Li, X., Chen, B., Fan, J., Kang, J., Ye, J., Wang, X., & Niyato, D. (2024). Cloud-edge-end collaborative intelligent service computation offloading: A digital twin driven edge coalition approach for industrial IoT. *IEEE Transactions on Network and Service Management*, *21*(6), 6318-6330.

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(4), e1507.

Oloruntoba, O. (2024). Green cloud computing: AI for sustainable database management. *World Journal of Advanced Research and Reviews*, *23*(03), 3242-3257.