**Technical Requirements**

**Document Version:** 1.0  
**Last Updated:** September 2025  
**Author:** Dr. Yishuang Xu, University of Manchester

**Executive Summary**

This document specifies the technical requirements for implementing the AI-Powered Sustainability Regulatory Compliance Platform. It defines the technology stack, infrastructure requirements, AI/ML specifications, and integration protocols necessary to deliver the functional requirements outlined in FR-001.

**System Architecture Requirements**

**Technology Stack**

**Backend Services**

* **Programming Language:** Python 3.9+
* **Web Framework:** FastAPI 0.104+
* **Database:** PostgreSQL 15+ (primary), Redis 7+ (caching)
* **Message Queue:** Apache Kafka 3.5+ or RabbitMQ 3.12+
* **Task Queue:** Celery 5.3+ with Redis broker

**AI/ML Stack**

* **Deep Learning:** PyTorch 2.1+ or TensorFlow 2.14+
* **NLP:** spaCy 3.7+, Transformers 4.35+, OpenAI GPT-4 API
* **Computer Vision:** OpenCV 4.8+, PIL, pdf2image
* **Time Series:** Prophet, statsmodels, scikit-learn 1.3+
* **Graph Networks:** PyTorch Geometric, NetworkX

**Frontend**

* **Framework:** React 18+ with TypeScript 5+
* **UI Library:** Material-UI 5+ or Ant Design 5+
* **Data Visualization:** D3.js 7+, Recharts 2.8+, Plotly.js
* **State Management:** Redux Toolkit or Zustand

**Infrastructure**

* **Containerization:** Docker 24+, Docker Compose
* **Orchestration:** Kubernetes 1.28+ (production)
* **Cloud Platform:** AWS, Azure, or Google Cloud Platform
* **CI/CD:** GitHub Actions, GitLab CI, or Jenkins
* **Monitoring:** Prometheus, Grafana, ELK Stack

**AI/ML Component Specifications**

**Natural Language Processing Module**

**Regulatory Text Processing**

Component: EU Taxonomy Text Analyzer

Purpose: Parse and interpret 24,000+ pages of EU regulatory documents

Requirements:

- Language Models: BERT-based transformer (DistilBERT-multilingual)

- Document Processing: 1000+ pages/hour processing speed

- Accuracy: 95%+ entity extraction accuracy

- Languages: English, German, French, Spanish, Italian

- Memory: Maximum 8GB RAM per processing instance

- Storage: 50GB for pre-trained models and regulatory corpus

**Implementation Requirements:**

* **Pre-processing Pipeline:** Clean regulatory PDFs, extract structured text, normalize formatting
* **Named Entity Recognition:** Extract regulatory entities (article numbers, technical criteria, thresholds)
* **Semantic Similarity:** Compare property descriptions against technical screening criteria
* **Change Detection:** Identify modifications in regulatory amendments using diff algorithms

**Computer Vision Module**

**Document Processing Engine**

Component: Property Document Analyzer

Purpose: Extract data from utility bills, EPCs, and building certificates

Requirements:

- OCR Engine: Tesseract 5+ with custom training data

- Image Processing: OpenCV for preprocessing and enhancement

- Accuracy: 95%+ for printed text, 85%+ for handwritten text

- Formats: PDF, JPEG, PNG, TIFF (up to 300 DPI)

- Processing Speed: 100 documents/hour per instance

- GPU: Optional NVIDIA GPU for faster processing

**Implementation Requirements:**

* **Image Preprocessing:** Deskew, denoise, enhance contrast, remove artifacts
* **Template Matching:** Recognize common utility bill formats and layouts
* **Data Validation:** Cross-reference extracted values with expected ranges
* **Quality Scoring:** Assign confidence scores to extracted data points

**Time Series Forecasting Module**

**Energy Consumption Predictor**

Component: Energy Forecasting Engine

Purpose: Predict future energy consumption and carbon emissions

Requirements:

- Models: LSTM, Prophet, ARIMA ensemble

- Accuracy: <10% MAPE for 12-month forecasts

- Features: Weather data, occupancy patterns, building characteristics

- Training Data: Minimum 24 months historical data

- Inference Time: <5 seconds per property

- Model Update: Monthly retraining capability

**Implementation Requirements:**

* **Feature Engineering:** Seasonal decomposition, lag features, external variables
* **Model Selection:** Automated hyperparameter tuning with cross-validation
* **Uncertainty Quantification:** Prediction intervals and confidence bounds
* **Drift Detection:** Monitor model performance and trigger retraining

**Graph Neural Network Module**

**Portfolio Relationship Mapper**

Component: Property Portfolio Analyzer

Purpose: Map relationships and optimize portfolio performance

Requirements:

- Framework: PyTorch Geometric or DGL

- Graph Size: Up to 10,000 nodes (properties)

- Edge Types: Geographic, performance, ownership relationships

- Analysis: Community detection, centrality measures, optimization

- Memory: Scalable to 32GB+ for large portfolios

**Database Requirements**

**Primary Database (PostgreSQL)**

**Schema Design**

-- Core Tables

properties (id, name, address, property\_type, size, acquisition\_date)

sustainability\_metrics (property\_id, metric\_type, value, date, source)

compliance\_assessments (property\_id, regulation\_type, status, score, date)

regulatory\_updates (id, regulation, change\_type, effective\_date, content)

documents (id, property\_id, document\_type, file\_path, processed\_status)

-- AI Results

ml\_predictions (property\_id, prediction\_type, value, confidence, date)

document\_extractions (document\_id, field\_name, value, confidence)

taxonomy\_mappings (property\_id, activity\_code, alignment\_score, evidence)

**Performance Requirements**

* **Concurrent Users:** 100+ simultaneous connections
* **Query Performance:** <1 second for dashboard queries
* **Data Retention:** 10+ years of historical data
* **Backup:** Daily automated backups with point-in-time recovery
* **Indexing:** Optimized indexes for time-series and geospatial queries

**Cache Layer (Redis)**

**Caching Strategy**

* **Session Data:** User sessions and authentication tokens
* **API Responses:** Frequently accessed compliance data (TTL: 1 hour)
* **ML Model Cache:** Prediction results (TTL: 24 hours)
* **Document Processing:** OCR results (TTL: 7 days)

**API Specifications**

**RESTful API Requirements**

**Authentication & Authorization**

Authentication: JWT tokens with refresh mechanism

Authorization: Role-based access control (RBAC)

Rate Limiting: 1000 requests/hour per user, 10000/hour per organization

API Versioning: Semantic versioning (v1, v2, etc.)

Documentation: OpenAPI 3.0 specification with Swagger UI

**Core Endpoints**

GET /api/v1/properties

- Purpose: List properties with filtering and pagination

- Response Time: <500ms

- Caching: 15 minutes

POST /api/v1/properties/{id}/documents

- Purpose: Upload and process property documents

- File Size Limit: 50MB per file

- Supported Formats: PDF, JPEG, PNG

- Processing: Asynchronous with status tracking

GET /api/v1/compliance/{property\_id}/taxonomy

- Purpose: Get EU Taxonomy alignment assessment

- Response Time: <1 second

- Caching: 1 hour

POST /api/v1/predictions/energy

- Purpose: Generate energy consumption forecasts

- Response Time: <5 seconds

- Input: Property ID, forecast period

**Webhook System**

Events: document\_processed, compliance\_updated, regulatory\_change

Delivery: HTTP POST with retry mechanism (exponential backoff)

Security: HMAC-SHA256 signature verification

Timeout: 30 seconds per webhook call

**Integration Requirements**

**Property Management Systems**

**Integration Protocols**

Yardi Voyager:

- Protocol: REST API or SOAP

- Authentication: OAuth 2.0 or API key

- Data Format: JSON or XML

- Sync Frequency: Daily batch or real-time webhooks

MRI Software:

- Protocol: REST API

- Authentication: API key with tenant isolation

- Data Format: JSON

- Rate Limits: 1000 calls/hour

**Data Mapping Requirements**

* **Property Master Data:** Address, type, size, ownership details
* **Lease Information:** Tenant details, lease terms, occupancy rates
* **Utility Data:** Consumption readings, billing periods, costs
* **Maintenance Records:** Equipment installations, upgrades, certifications

**External Data Sources**

**Weather Data Integration**

Providers: OpenWeatherMap, Weather Underground, ECMWF

Data Points: Temperature, humidity, solar radiation, heating/cooling degree days

Frequency: Hourly updates

Storage: 2+ years of historical data

API Limits: Respect provider rate limits and quotas

**Regulatory Database Integration**

EU Official Databases:

- EUR-Lex: Legal document updates

- ESMA: Technical standards and guidelines

- EBA: Banking regulation updates

Monitoring: Daily automated checks for new publications

Alert Mechanism: Real-time notifications for relevant changes

**Security Requirements**

**Data Protection**

**Encryption**

In Transit: TLS 1.3 for all API communications

At Rest: AES-256 encryption for database and file storage

Key Management: AWS KMS, Azure Key Vault, or HashiCorp Vault

Certificate Management: Automated SSL certificate renewal

**Access Control**

Authentication: Multi-factor authentication (MFA) required

Password Policy: Minimum 12 characters, complexity requirements

Session Management: Secure session tokens with expiration

Audit Logging: All data access and modifications logged

**Compliance Requirements**

**GDPR Compliance**

* **Data Minimization:** Collect only necessary personal data
* **Right to Erasure:** Implement data deletion mechanisms
* **Data Portability:** Export user data in standard formats
* **Consent Management:** Track and manage user consent preferences

**SOC 2 Type II Requirements**

* **Security:** Implement comprehensive security controls
* **Availability:** 99.9% uptime SLA with monitoring
* **Processing Integrity:** Ensure data accuracy and completeness
* **Confidentiality:** Protect confidential information
* **Privacy:** Implement privacy protection measures

**Performance Requirements**

**Scalability Specifications**

**Horizontal Scaling**

Application Servers: Auto-scaling from 2 to 20 instances

Database: Read replicas for query distribution

Queue Workers: Elastic scaling based on queue depth

Load Balancer: Support for 10,000+ concurrent connections

**Performance Benchmarks**

API Response Times:

- Simple queries: <500ms (95th percentile)

- Complex analytics: <5 seconds

- Document processing: <30 seconds per document

- Report generation: <2 minutes for 1000 properties

Database Performance:

- Simple queries: <100ms

- Complex aggregations: <2 seconds

- Bulk operations: 10,000 records/second

**Resource Requirements**

**Production Environment**

Application Servers:

- CPU: 8+ cores per instance

- RAM: 16GB+ per instance

- Storage: 100GB+ SSD per instance

Database Server:

- CPU: 16+ cores

- RAM: 64GB+

- Storage: 1TB+ SSD with RAID 10

ML Processing:

- CPU: 16+ cores or GPU (NVIDIA V100/A100)

- RAM: 32GB+

- Storage: 500GB+ for models and temporary data

**Development Requirements**

**Code Quality Standards**

**Python Code Standards**

Style Guide: PEP 8 with Black formatter

Type Hints: Required for all function signatures

Documentation: Docstrings in Google or NumPy format

Testing: 90%+ code coverage with pytest

Linting: pylint, flake8, mypy for static analysis

**JavaScript/TypeScript Standards**

Style Guide: ESLint with Airbnb configuration

Type Safety: Strict TypeScript configuration

Testing: Jest with React Testing Library

Code Coverage: 85%+ coverage requirement

**CI/CD Pipeline Requirements**

**Automated Testing**

Unit Tests: Run on every commit

Integration Tests: Run on pull requests

End-to-End Tests: Run on staging deployment

Performance Tests: Weekly automated runs

Security Scans: SAST and DAST tools integration

**Deployment Pipeline**

Environments: Development, Staging, Production

Deployment Strategy: Blue-green or rolling deployment

Rollback Capability: Automated rollback on failure detection

Infrastructure as Code: Terraform or CloudFormation

Secrets Management: Secure handling of API keys and credentials

**Monitoring and Observability**

**Application Monitoring**

**Metrics Collection**

Application Metrics: Response times, error rates, throughput

Business Metrics: Processing volumes, compliance scores, user activity

Infrastructure Metrics: CPU, memory, disk, network utilization

Custom Metrics: AI model performance, data quality scores

**Alerting Rules**

Critical Alerts: System downtime, database failures, security breaches

Warning Alerts: High response times, resource utilization >80%

Business Alerts: Failed compliance calculations, regulatory changes

Escalation: PagerDuty or similar for 24/7 critical alerts

**Logging Requirements**

**Log Management**

Structure: JSON format for structured logging

Levels: DEBUG, INFO, WARNING, ERROR, CRITICAL

Retention: 90 days for application logs, 7 years for audit logs

Centralization: ELK Stack or equivalent for log aggregation

Search: Full-text search and advanced filtering capabilities

**Disaster Recovery**

**Backup Strategy**

Database Backups: Daily full backups, hourly incremental

File Storage: Daily backups with version control

Configuration: Infrastructure configuration in version control

Recovery Time Objective (RTO): 4 hours maximum

Recovery Point Objective (RPO): 1 hour maximum

**Business Continuity**

Multi-Region Setup: Primary and secondary regions

Failover Mechanism: Automated failover for critical components

Data Replication: Real-time or near-real-time data synchronization

Testing: Quarterly disaster recovery drills

Documentation: Detailed runbooks for emergency procedures

**Document Control:**

* Document ID: TR-001
* Version: 1.0
* Review Date: December 2025
* Dependencies: FR-001 (Functional Requirements)
* Approval: Innovation Factory, University of Manchester