Методологія підготовки даних до аналізу в інтересах інформаційно-аналітичних систем об'єднаного угруповання військ

Розділ 1: Фундаментальна роль підготовки даних у досягненні переваги в ухваленні рішень

1.1. Переосмислення підготовки даних: від технічного етапу до стратегічного імперативу

У сучасній парадигмі ведення бойових дій, що дедалі більше спирається на дані, перевага в ухваленні рішень є ключовим фактором досягнення перемоги. Ця перевага досягається не лише завдяки обсягу зібраної інформації, а й завдяки швидкості та якості її перетворення на дієві розвідувальні дані та обґрунтовані команди. У цьому контексті підготовка даних перестає бути суто технічним, підготовчим етапом і трансформується у стратегічний імператив, що безпосередньо впливає на боєздатність та ефективність об'єднаного угруповання військ (ОУВ).

Традиційно в галузі науки про дані домінує так зване "правило 80/20", згідно з яким аналітики витрачають приблизно 80% свого часу на збір, очищення та реорганізацію даних і лише 20% – на власне аналіз та моделювання. У комерційному секторі ці 80% є операційними витратами, що впливають на рентабельність. Однак у військовому контексті цей часовий лаг є критичною вразливістю. Це "вікно у 80%" – це час, протягом якого супротивник може діяти, змінювати диспозицію та захоплювати ініціативу, поки наші аналітичні системи ще не готові надати релевантну інформацію.

Ця проблема напряму пов'язана з концепцією циклу СОРУ (Спостереження – Орієнтація – Рішення – Ухвалення), розробленою полковником Джоном Бойдом. Цикл СОРУ стверджує, що сторона, яка здатна проходити цей цикл швидше за супротивника, отримує вирішальну тактичну та стратегічну перевагу. Сучасні інформаційно-аналітичні системи (ІАС) є ядром фаз "Орієнтація" та "Рішення". Якщо 80% часу, відведеного на ці фази, витрачається на підготовку даних, це штучно та передбачувано подовжує наш цикл СОРУ, віддаючи ініціативу ворогу.

Отже, фундаментальна мета методології підготовки даних для ІАС ОУВ полягає не просто в забезпеченні якості даних для моделей, а в систематичному скороченні цього "80%-го вікна". Це вимагає переходу від ручних, разових процесів до індустріалізованого, автоматизованого та керованого підходу. Інвестиції в автоматизацію очищення даних, створення надійних конвеєрів обробки (pipelines) та впровадження суворої системи управління даними (Data Governance) є прямими інвестиціями у скорочення циклу СОРУ. Це перетворює підготовку даних із допоміжної функції на бойовий мультиплікатор, що підвищує оперативний темп та дозволяє випереджати дії супротивника. Фокус зміщується з питання "як очистити дані?" на стратегічне завдання "як створити сили, що завжди готові до роботи з даними?".

1.2. Унікальні виклики військових даних в середовищі ІАЗ ОУВ

Дані, що циркулюють в інформаційному просторі ІАС ОУВ, кардинально відрізняються від даних у корпоративному чи академічному середовищі. Якщо класична концепція "великих даних" описується чотирма 'V' (Volume – обсяг, Velocity – швидкість, Variety – різноманітність, Veracity – достовірність), то у військовому контексті кожна з цих характеристик набуває екстремального та загрозливого характеру.

- **Різноманітність (Variety):** ІАС ОУВ змушена обробляти надзвичайно гетерогенні потоки даних. Це включає структуровані дані (логістичні таблиці, бази даних особового складу), напівструктуровані дані (оперативні зведення, текстові доповіді у формалізованому вигляді) та неструктуровані дані, які складають левову частку інформації (супутникові знімки та відео з БПЛА (IMINT), радіоперехоплення (SIGINT), доповіді агентурної розвідки (HUMINT), інформація з відкритих джерел (OSINT), акустичні дані, сейсмічні показники). Інтеграція цих різнорідних форматів є нетривіальним технічним завданням.
- **Достовірність (Veracity):** Це найкритичніша відмінність військових даних. У комерційному світі проблеми з достовірністю даних (Veracity) зазвичай є наслідком випадкових помилок: збоїв сенсорів, помилок при ручному введенні, системних збоїв. У військовому середовищі до цього додається п'яте, визначальне 'V' – **Vulnerability (вразливість)** до цілеспрямованих дій супротивника. Помилки та

аномалії в даних можуть бути не випадковим шумом, а результатом активної ворожої протидії: радіоелектронної боротьби (РЕБ), GPS-спуфінгу, поширення дезінформації, кібератак на сенсори та канали зв'язку, використання маскувальних засобів для спотворення даних візуальної розвідки.

Ця обставина докорінно змінює підхід до забезпечення якості даних. Стандартні протоколи очищення даних, розроблені для виправлення стохастичних (випадкових) помилок, стають не просто недостатніми, а й небезпечними. Наприклад, стандартний алгоритм виявлення аномалій може ідентифікувати нетипове значення в даних сенсора як "викид" і видалити його. Однак у бойовій обстановці цей "викид" може бути сигналом роботи ворожої системи РЕБ. Прогалини в потоці даних можуть бути не технічним збоєм, а свідченням знищення вузла зв'язку. Застосування стандартних методів очищення, таких як видалення викидів або заповнення пропусків середнім значенням, може ненавмисно знищити безцінну розвідувальну ознаку.

Таким чином, методологія підготовки даних для ІАС ОУВ повинна розглядати забезпечення якості даних як невід'ємну частину контррозвідувальної діяльності. Необхідно впроваджувати подвійний підхід до обробки аномалій. Конвеєр даних повинен мати один процес для ідентифікації та корекції очікуваних випадкових помилок (наприклад, калібрування сенсорів). Паралельно має діяти другий процес, який прапорує, ізолює та негайно ескалує підозрілі, нетипові дані для поглибленого аналізу фахівцями з розвідки та кібербезпеки. Це перетворює команду підготовки даних з технічних фахівців на першу лінію оборони в інформаційній війні, яка здатна виявляти приховані дії супротивника на найнайранішому етапі – на рівні сирих даних.

Розділ 2: Архітектура конвеєра підготовки даних для критично важливого моделювання

2.1. Організаційна структура та управління даними (Data Governance)

Ефективна підготовка даних неможлива без чіткої організаційної структури та формалізованих процесів управління. У військовому середовищі, де ціна помилки вимірюється життями та стратегічними втратами, впровадження програми управління даними (Data Governance) є не бюрократичною формальністю, а фундаментальною

вимогою для забезпечення надійності та підзвітності всього аналітичного процесу.

Центральним елементом такої програми є визначення чітких ролей та відповідальностей. Ключовими фігурами є:

- **Розпорядники даних (Data Stewards):** Це не IT-фахівці, а експерти у предметній області, які несуть відповідальність за якість, цілісність та безпеку даних у своєму домені. Наприклад, офіцер управління логістики (G4) є природним розпорядником для даних про постачання пального, боєприпасів та продовольства. Офіцер розвідки (G2) – розпорядник для розвідувальних зведень. Їхнє завдання – визначати правила якості даних, затверджувати стандарти та виступати кінцевим арбітром у вирішенні питань щодо інтерпретації та коректності даних у своїй зоні відповідальності.
- **Аналітики якості даних (Data Quality Analysts):** Це технічні спеціалісти, які реалізують політики, визначені розпорядниками даних, розробляють та впроваджують автоматизовані перевірки якості, моніторять потоки даних на предмет аномалій та звітують про стан якості даних.

Ця структура відображає фундаментальний принцип військового управління – єдиноначальність та персональну відповідальність – і поширює його на інформаційний домен. Командир несе відповідальність за рішення, ухвалені на основі даних, наданих IAC. Щоб ця відповідальність була обґрунтованою, командир повинен мати впевненість в інформаційному ланцюгу постачання. Програма управління даними створює формальний механізм для цієї впевненості. Роль розпорядника даних аналогічна ролі офіцера, відповідального за матеріальні активи, такі як зброя чи техніка. Дані стають таким самим підзвітним ресурсом.

Невід'ємним технічним компонентом управління даними є **відстеження походження даних (data lineage)**. Це процес документування повного життєвого циклу даних: звідки вони надійшли, які трансформації до них застосовувалися, ким і коли. У разі ухвалення критичного рішення, яке призвело до негативних наслідків, відстеження походження даних дозволяє провести швидкий та точний "розбір польотів", встановити, на якому етапі виникла помилка – чи були дані невірними на вході, чи їх неправильно трансформували, чи модель їх невірно інтерпретувала. Без такого аудиторського сліду неможливо ані встановити причину збою, ані запобігти його повторенню. Таким чином, управління даними та відстеження їх походження є прямою реалізацією принципу відповідальності командування в епоху цифрової війни.

2.2. Технічна архітектура: від ETL до автоматизованих конвеєрів

Технічною основою процесу підготовки даних є конвеєри (pipelines), які автоматизують переміщення та обробку даних від джерела до аналітичної системи. Історично домінував

підхід **ETL (Extract, Transform, Load – Видобування, Трансформація, Завантаження)**. У цьому процесі дані спочатку видобуваються з джерела, потім трансформуються згідно з наперед визначеною схемою та бізнес-правилами, і лише після цього завантажуються в кінцеве сховище (зазвичай, структуроване сховище даних). Цей підхід забезпечує високий рівень контролю та гарантує, що в сховище потрапляють лише чисті, стандартизовані дані. Однак він є повільним та негнучким. Якщо в майбутньому виникає нова аналітична потреба, яка вимагає доступу до сирих даних, що були відфільтровані на етапі трансформації, їх доводиться видобувати заново.

Сучаснішою альтернативою є архітектура **ELT (Extract, Load, Transform – Видобування, Завантаження, Трансформація)**. У цьому підході сирі, необроблені дані спочатку видобуваються і негайно завантажуються у централізоване, гнучке сховище, яке називається "озером даних" (data lake). Трансформації застосовуються вже після завантаження, безпосередньо перед аналізом, відповідно до конкретних потреб завдання. Це забезпечує надзвичайну гнучкість та швидкість, оскільки всі дані, в їхньому первісному вигляді, завжди доступні для аналізу.

Для ІАС ОУВ вибір між ETL та ELT є стратегічним рішенням. У військовому контексті концепцію "озера даних" слід переосмислити як **"стратегічний резерв даних"**. Рішення про впровадження архітектури ELT – це не просто технічний вибір, а свідоме рішення командування зберігати кожен фрагмент зібраної інформації в її первозданному вигляді. Причина цього проста: вимоги до розвідки та аналізу є динамічними і постійно змінюються. Дані, які сьогодні здаються нерелевантними і були б відфільтровані жорстким ETL-процесом (наприклад, фонові атмосферні шуми, дані про цивільний трафік), завтра можуть стати ключовими для відповіді на нове розвідувальне питання (наприклад, аналіз впливу погоди на роботу нової ворожої системи РЕБ або виявлення аномалій у цивільному трафіку, що свідчать про приховану підготовку супротивника).

Зберігаючи сирі дані у стратегічному резерві, ОУВ забезпечує собі аналітичну гнучкість. Це дозволяє "повернутися в часі" і повторно проаналізувати історичні дані крізь призму нових загроз, нових розвідувальних даних та нових аналітичних методів. Такий підхід максимізує довгострокову цінність кожного зібраного байта інформації, перетворюючи дані з витратного матеріалу для поточних завдань на довгостроковий стратегічний актив.

2.3. Людина-в-циклі (Human-in-the-Loop): Інтеграція експертних знань

Незважаючи на стрімкий розвиток автоматизації та штучного інтелекту, людський досвід та інтуїція залишаються незамінними, особливо у складних та неоднозначних військових сценаріях. Автоматизація – це не заміна експерта, а інструмент для підвищення його

ефективності. Архітектура підготовки даних для ІАС ОУВ повинна бути побудована на принципі **"людина-в-циклі" (Human-in-the-Loop, HITL)**.

У системі HITL автоматизовані процеси виконують переважну більшість рутинних завдань: перевірку форматів, стандартизацію одиниць виміру, виявлення очевидних помилок. Однак, коли система стикається з неоднозначною, аномальною або потенційно критичною інформацією, вона не ухвалює остаточного рішення, а автоматично створює завдання для людини-експерта. Наприклад:

- Алгоритм комп'ютерного зору може ідентифікувати колону техніки на супутниковому знімку, але позначити її як "невідомий тип". Знімок автоматично направляється аналітику візуальної розвідки для ідентифікації.
- Система обробки природної мови може виявити в радіоперехопленні нове кодове слово. Фрагмент перехоплення направляється лінгвісту для аналізу.
- Моніторинг логістичних даних може зафіксувати різке, нетипове падіння запасів пального у підрозділі. Система створює сповіщення для офіцера-логіста для перевірки можливого витоку, крадіжки або неврахованих витрат.

Однак справжня сила підходу HITL полягає не лише у валідації даних. Кожна взаємодія експерта з системою є безцінною можливістю для її навчання. Коли аналітик ідентифікує техніку на знімку, його дія (наприклад, присвоєння мітки "танк Т-72Б3") разом із самим зображенням повинна зберігатися як новий, високоякісний приклад для навчання моделі комп'ютерного зору. Коли лінгвіст розшифровує нове кодове слово, це поповнює словник системи обробки мови.

Таким чином, створюється безперервний цикл зворотного зв'язку: система автоматизує рутину, експерт вирішує складні випадки, а рішення експерта використовуються для донавчання системи. З часом система стає "розумнішою", вона вбирає в себе приховані, неформалізовані знання своїх експертних користувачів. Кількість винятків, що потребують ручного втручання, зменшується, а точність автоматичної обробки зростає. У цій парадигмі експерт еволюціонує від ролі "очищувача даних" до ролі "вчителя" для штучного інтелекту, що дозволяє масштабувати унікальні людські знання на весь обсяг вхідної інформації.

Розділ 3: Огляд сучасних методів підготовки даних та їх застосування

3.1. Фундаментальні техніки: очищення, імпутація та нормалізація

Основою будь-якого конвеєра підготовки даних є набір фундаментальних технік, спрямованих на виправлення помилок, заповнення пропусків та приведення даних до єдиного формату. Вибір конкретного методу залежить від типу даних, їхнього походження та потенційного впливу на кінцеву аналітичну модель.

Очищення даних включає виявлення та виправлення помилок. Це може бути валідація форматів (наприклад, перевірка, що координати знаходяться у допустимому діапазоні), виявлення дублікатів, а також ідентифікація аномалій або "викидів". Як зазначалося раніше, у військовому контексті аномалії вимагають особливої уваги, оскільки можуть бути сигналом ворожої активності.

Імпутація пропущених даних – це процес заповнення відсутніх значень. Пропуски в даних є неминучими (збій сенсора, втрата зв'язку, неповний звіт), але їхня неправильна обробка може суттєво спотворити результати аналізу. Наприклад, використання простого середнього значення для заповнення пропуску в даних про витрату пального бронетанковим підрозділом може небезпечно занизити реальну потребу, якщо середнє розраховано по всіх типах підрозділів, включаючи немеханізовані. Вибір методу імпутації є критично важливим.

Нижче наведено порівняльний аналіз поширених методів імпутації в контексті їхньої придатності для ІАС ОУВ.

Таблиця 3.1: Порівняльний аналіз методів імпутації пропущених даних

Метод	Опис	Обчислювальна складність	Припущення	Придатність для даних ОУВ	Ключові обмеження та вразливості до дій супротивника
Імпутація середнім/ медіаною/ модою	Заміна пропущених значень середнім арифметичним, медіаною	Дуже низька	Дані відсутні абсолютно випадково (MCAR). Немає зв'язку між	Обмежена. Може використовуватись для некритичних даних	**Висока вразливість.** Супротивник, знаючи про використа

	або найчастіши м значенням по всьому стовпцю.		відсутністю даних та іншими змінними.	або як початковий етап.	ння цього методу, може заглушити сенсор (створивш и пропуски), щоб штучно занизити або завищити середнє значення, що призведе до невірних прогнозів.
Регресійн а імпутація	Побудова регресійно ї моделі для прогнозува ння пропущено го значення на основі інших змінних у тому ж записі.	Середня	Існує лінійна залежність між змінними.	Середня. Корисна для даних, де є сильні кореляції (напр., залежність витрати пального від пройденої відстані та типу місцевості) .	Може генерувати нереалісти чні прогнози, якщо залежності нелінійні. Вразлива, якщо супротивн ик може впливати на предиктор и (напр., спотворюв ати дані про відстань).

Імпутація методом K-найближчих сусідів (KNN)	Пошук k найближчих (найбільш схожих) записів у наборі даних, де значення не пропущене, та використання їхніх значень (напр., усереднення) для заповнення пропуску.	Висока	Схожі об'єкти (напр., підрозділи) мають схожі характеристики.	Висока. Дуже ефективна для даних про підрозділи, техніку, логістику. Наприклад, для оцінки боєзапасу підрозділу можна знайти схожі підрозділи за типом, чисельністю та інтенсивністю бойових дій.	Вимагає ретельного вибору метрики відстані. Може бути чутливою до "прокляття розмірності", якщо ознак дуже багато.
Множинна імпутація (MICE)	Створення декількох (m > 1) повних наборів даних шляхом ітеративного заповнення пропусків на основі розподілу даних. Аналіз проводиться на кожному	Дуже висока	Дані відсутні випадково (MAR). Відсутність може залежати від спостережуваних даних, але не від неспостережуваних.	Дуже висока. Вважається золотим стандартом, оскільки враховує невизначеність, пов'язану з імпутацією. Ідеальна для критично важливих моделей, де потрібна	Складна в реалізації та інтерпретації. Вимагає значних обчислювальних ресурсів.

	наборі, а результати усереднюються.			максимальна точність.	

Нормалізація та стандартизація – це процес приведення даних до єдиного масштабу. Це необхідно для багатьох алгоритмів машинного навчання, які чутливі до масштабу ознак (наприклад, методи кластеризації, машини опорних векторів). Стандартизація також включає приведення даних до єдиних форматів: перетворення всіх географічних координат у єдину систему (напр., WGS 84), уніфікація часових поясів (зазвичай, UTC), переведення всіх одиниць виміру в систему CI (напр., відстані в кілометрах, вага в кілограмах). Це усуває неоднозначність та забезпечує коректність подальших обчислень.

3.2. Інтеграція даних та розв’язання сутностей (Entity Resolution)

Одним із найскладніших, але й найважливіших завдань у підготовці даних для ІАС ОУВ є злиття інформації з різнорідних джерел для створення єдиної, цілісної оперативної картини. Цей процес, відомий як **розв'язання сутностей (entity resolution)**, полягає в ідентифікації та об'єднанні записів, що стосуються однієї й тієї ж реальної сутності (наприклад, ворожого підрозділу, одиниці техніки, особи) у різних наборах даних.

Наприклад, ІАС може отримати три окремі повідомлення:

1. **SIGINT:** Радіоперехоплення, в якому згадується умовне найменування "Підрозділ 734", що рухається на схід.
2. **IMINT:** Знімок з БПЛА, що показує колону з 10 танків та 3 БМП за координатами 48.123,37.456.
3. **HUMINT:** Доповідь агента про "мотострілецький батальйон", що посилюється танками і готується до наступу в певному районі.

Завдання розв'язання сутностей – визначити, чи всі ці повідомлення стосуються однієї й тієї ж загрози. У комерційному секторі метою цього процесу є створення так званого "золотого запису" (golden record) – єдиного, чистого профілю клієнта, де всі суперечливі дані усуваються на користь найбільш достовірних.

Однак в розвідувальному аналізі такий підхід є небезпечним спрощенням. Суперечлива інформація – це не "шум", який потрібно усунути, а життєво важливий сигнал. Розбіжність між двома джерелами може вказувати на помилку одного з них, на активну дезінформаційну кампанію супротивника або на реальну зміну обстановки. Тому мета розв'язання сутностей в ІАС ОУВ полягає не у створенні "золотого запису", а у

формуванні **"живого досьє" (living file)** на кожну сутність.

Таке "досьє" не просто об'єднує дані, а й моделює зв'язки між ними. Воно зберігає походження кожного фрагмента інформації та явно фіксує відносини:

- **Підтвердження:** Джерело А підтверджує інформацію з джерела Б.
- **Суперечність:** Джерело В надає інформацію, що суперечить даним з джерел А і Б.
- **Невизначеність:** Недостатньо даних для однозначного зв'язку повідомлення Г з існуючою сутністю.

Кінцевим продуктом є не просто твердження "Ворожий батальйон знаходиться в точці Y", а значно багатший аналітичний висновок: "Джерела SIGINT та IMINT з високою ймовірністю вказують на знаходження 1-го мотострілецького батальйону 5-ї бригади в районі Y. Водночас, джерело HUMINT, яке раніше демонструвало середню надійність, повідомляє про його переміщення в район Z. Ця розбіжність може свідчити про підготовку до наступу або про дезінформацію". Такий нюансований, заснований на доказах висновок є набагато цінішим для командира, ніж оманливо простий "золотий запис", оскільки він відображає реальну невизначеність та ризики бойової обстановки.

3.3. Просунута інженерія ознак (Feature Engineering) для прогнозної аналітики

Якщо очищення та інтеграція готують дані до використання, то **інженерія ознак (feature engineering)** перетворює їх на потужний інструмент для прогнозного моделювання. Це процес створення нових, інформативних змінних (ознак) з наявних сирих даних. Саме на цьому етапі відбувається найглибша синергія між машиною та людиною, оскільки інженерія ознак дозволяє "закодувати" досвід, знання та інтуїцію військового експерта у формат, зрозумілий для алгоритмів машинного навчання.

Машина може знайти статистичні кореляції в будь-яких даних, але без правильних ознак ці кореляції можуть бути поверхневими або навіть хибними. Військовий аналітик, на відміну від машини, розуміє, що важливою є не просто серія GPS-координат ворожої одиниці, а швидкість її переміщення, напрямок руху, час доби, близькість до основних доріг, укриттів чи відомих позицій. Створення ознак, що відображають ці концепції, є суттю інженерії ознак.

Приклади інженерії ознак у військовому контексті:

- **Часові ознаки (Temporal):** На основі періодичних логістичних звітів про залишки боєприпасів можна створити ознаку rate_of_consumption (темп витрати) за останню добу. Високе значення цієї ознаки може бути потужним предиктором підготовки до

наступальних дій.

- **Просторові ознаки (Geospatial):** Маючи координати ворожих та дружніх підрозділів, можна обчислити низку критично важливих ознак: distance_to_nearest_friendly_unit (відстань до найближчого дружнього підрозділу), distance_to_critical_infrastructure (відстань до об'єкта критичної інфраструктури), time_to_reach_objective (розрахунковий час досягнення цілі з поточною швидкістю). Ці ознаки є набагато інформативнішими для моделі оцінки загроз, ніж сирі координати.
- **Текстові ознаки (NLP):** За допомогою методів обробки природної мови (Natural Language Processing, NLP) з текстових зведень та перехоплень можна видобувати структуровану інформацію: іменовані сутності (імена командирів, номери підрозділів, географічні назви), ключові дії (наступ, відступ, перегрупування), а також аналізувати тональність (sentiment analysis) комунікацій, що може вказувати на падіння чи зростання морального духу супротивника.

Таким чином, інженерія ознак є процесом операціоналізації військової доктрини та тактичних знань. Коли аналітик створює ознаку is_in_defensive_formation (чи знаходиться в оборонному шикуванні) або violates_communication_protocol (чи порушує протокол зв'язку), він фактично перекладає свій бойовий досвід мовою математики. Це дозволяє моделям машинного навчання вчитися не на абстрактних статистичних патернах, а на ознаках, що відображають фундаментальні принципи ведення війни. Якість інженерії ознак безпосередньо залежить від глибини експертних знань аналітика і є найпотужнішим важелем для підвищення точності та релевантності будь-якої прогнозної моделі в ІАС ОУВ.

Розділ 4: Ілюстративні сценарії та найкращі практики в контексті ОУВ

Теоретичні концепції підготовки даних набувають реального значення лише тоді, коли вони застосовуються для вирішення конкретних бойових завдань. Розглянемо два типових сценарії, що ілюструють застосування описаної методології в роботі ІАС ОУВ.

4.1. Сценарій А: Підготовка даних для моделі прогнозування логістики

Проблема: Модель прогнозування потреб у пально-мастильних матеріалах (ПММ) для танкової бригади почала давати суттєво занижені результати, що призвело до критичної нестачі пального під час виконання бойового завдання. Аналіз показав, що модель не враховувала збільшення витрат пального при русі по пересіченій місцевості та ігнорувала неповні звіти від деяких підрозділів, що призвело до системної помилки.

Рішення з використанням методології підготовки даних:

1. **Збір та інтеграція даних:** Створюється єдиний набір даних, що об'єднує:
    - Структуровані дані з баз даних логістики: щоденні звіти про залишки ПММ, пробіг техніки, обсяги постачання.
    - Напівструктуровані дані: текстові запити на постачання від командирів підрозділів, що надходять через месенджери. За допомогою NLP з них витягуються ключові параметри (підрозділ, потреба, терміновість).
    - Геопросторові дані: треки руху техніки, карти місцевості з типами рельєфу (поля, ліси, болота).
    - Розвідувальні дані: прогнозована інтенсивність бойових дій на напрямку дії бригади.
2. **Управління та очищення даних:**
    - Призначається **розпорядник даних (Data Steward)** з управління логістики (G4), який відповідає за якість та повноту даних про ПММ.
    - Впроваджується автоматична перевірка: система прапорує будь-який звіт, де витрата пального на кілометр пробігу виходить за межі очікуваного діапазону для даного типу техніки.
    - Аномальні звіти (наприклад, нульова витрата при значному пробігу) направляються в систему **Human-in-the-Loop (HITL)** для перевірки відповідальним офіцером-логістом. Це дозволяє відрізнити реальну економію від помилки у звіті чи несправності датчика.
    - Пропущені звіти обробляються за допомогою **KNN-імпутації**: для підрозділу, що не надав звіт, система знаходить кілька найбільш схожих підрозділів (за типом техніки, інтенсивністю дій, типом місцевості) і розраховує ймовірну витрату на основі їхніх даних.
3. **Інженерія ознак:**
    - Створюється ключова ознака terrain_consumption_modifier (модифікатор витрати в залежності від місцевості), яка збільшує базову норму витрати при русі по бездоріжжю.
    - Розраховується динамічна ознака burn_rate_per_hour (витрата на годину) на основі даних про інтенсивність бойових дій.
    - Створюється прогнозна ознака days_of_supply_remaining (запас ходу в днях) для різних сценаріїв інтенсивності (низька, середня, висока).

Результат: На вхід моделі прогнозування подається чистий, повний та збагачений новими ознаками набір даних. Модель тепер враховує реальні умови експлуатації техніки та здатна надавати значно точніші прогнози потреб у ПММ, що дозволяє системі

логістики працювати на випередження та запобігати критичним збоям у постачанні.

4.2. Сценарій B: Злиття розвідданих з багатьох джерел для оцінки загрози

Проблема: Розвідувальні підрозділи отримують фрагментарну інформацію про активність супротивника з різних джерел. Існує високий ризик або недооцінити загрозу (якщо різні повідомлення про одну й ту ж ворожу групу будуть розглядатися окремо), або переоцінити її (якщо одне й те ж повідомлення, отримане різними каналами, буде враховано кілька разів).

Рішення з використанням методології підготовки даних:

1. **Збір та завантаження в "стратегічний резерв":**
   - Всі вхідні повідомлення – SIGINT, IMINT, HUMINT, OSINT – завантажуються в "озеро даних" у своєму первісному, сирому вигляді (архітектура ELT). Це зберігає весь контекст для майбутнього аналізу.
2. **Розв'язання сутностей та створення "живого досьє":**
   - Запускається процес **розв'язання сутностей**. Система використовує ймовірнісні моделі для зіставлення повідомлень за різними параметрами: час, місце, тип підрозділу, склад техніки, умовні найменування.
   - Наприклад, SIGINT-повідомлення про "Підрозділ 734" та IMINT-знімок колони танків зіставляються за часом та приблизним місцем розташування. Система створює нову сутність "Імовірна БТГр-1" і пов'язує з нею обидва повідомлення, вказуючи ймовірність зв'язку 85%.
   - Коли надходить HUMINT-звіт про "мотострілецький батальйон", система намагається зв'язати його з існуючою сутністю. Якщо дані (наприклад, склад техніки) суперечать попереднім, система не відкидає звіт, а додає його до "живого досьє" з позначкою "суперечлива інформація, потребує додаткової перевірки".
3. **Видобування та інженерія ознак:**
   - До кожного повідомлення в "досьє" застосовуються відповідні методи видобування ознак. Зі знімка IMINT витягуються кількість та типи техніки, аналізується її шикування. З текстового звіту HUMINT за допомогою NLP витягуються ключові дієслова ("готується", "атакує") та оцінюється рівень впевненості джерела.
   - Для сутності в цілому розраховуються узагальнюючі **геопросторові ознаки**: відстань до найближчого стратегічного об'єкта, приналежність до певної зони відповідальності, швидкість руху центру мас групи.

Результат: Замість розрізнених повідомлень аналітик отримує структуроване "живе досьє" на кожну потенційну загрозу. Це досьє містить всю пов'язану інформацію, включаючи суперечливу, з оцінкою достовірності та походженням кожного факту. На основі цього збагаченого набору даних модель класифікації загроз може працювати набагато точніше, оцінюючи рівень небезпеки кожної ідентифікованої групи супротивника та пріоритезуючи їх для подальшого спостереження чи вогневого ураження.

4.3. Рекомендації щодо інструментів, технологій та культури даних

Впровадження ефективної методології підготовки даних вимагає комплексного підходу, що охоплює технології, інструменти та, що найважливіше, організаційну культуру.

Технології та інструменти:

- **Баланс між Open-Source та COTS:** У військовому середовищі необхідно знаходити баланс між програмним забезпеченням з відкритим кодом (Python з бібліотеками Pandas, NumPy, Scikit-learn; Apache Spark для розподілених обчислень) та комерційними готовими рішеннями (COTS). Open-source надає гнучкість та можливість глибокої кастомізації, але вимагає високої кваліфікації персоналу та суворих процедур для забезпечення кібербезпеки. COTS-рішення можуть пропонувати інтегровані, сертифіковані для роботи в захищених мережах платформи, але можуть бути менш гнучкими та створювати залежність від постачальника. Вибір повинен ґрунтуватися на конкретних вимогах до безпеки, масштабованості та наявних компетенцій.
- **Платформи для спільної роботи:** Необхідно впроваджувати інструменти, що дозволяють аналітикам, розвідникам та фахівцям з даних спільно працювати над даними, ділитися ознаками, коментувати якість даних та відстежувати зміни. Це сприяє накопиченню колективних знань та прискорює аналітичний цикл.

Культура даних:

- **Дані як стратегічний актив:** Найважливішим елементом є зміна культури. Кожен військовослужбовець, від солдата, що вводить дані у тактичний планшет, до генерала, що ухвалює рішення на основі аналітичної панелі, повинен розуміти, що дані є таким же критично важливим ресурсом, як боєприпаси чи пальне. Це вимагає постійного навчання та роз'яснення.
- **Заохочення до співпраці:** Необхідно руйнувати "інформаційні бункери" між різними управліннями та родами військ. Ефективна підготовка даних можлива лише за умови тісної співпраці між операторами (джерелами даних), розвідниками (експертами з інтерпретації) та фахівцями з даних (технічними експертами).

- **Побудова довіри через прозорість:** Кінцевою метою всієї методології підготовки даних є побудова обґрунтованої **довіри** між людиною-командиром та машинною аналітичною системою. Ця довіра не може бути сліпою. Вона народжується з прозорості. Коли конвеєр підготовки даних є керованим, документованим та аудитованим, коли командир може бачити походження даних, на яких базується рекомендація системи, та розуміти ступінь їхньої надійності, він може ухвалити більш виважене та обґрунтоване рішення. Таким чином, прозорий та надійний процес підготовки даних не просто створює якісний набір даних; він генерує "оцінку впевненості" в самій інформації, що є найціннішим продуктом для ухвалення рішень в умовах високої невизначеності.